SEVEN AI LAWS

The Future of Mankind

A Techno-Legal-Philosophical Treatise on the Governance of Artificial Intelligence

By

Adv. Prashant Mali, Ph.D.

International Cyber Law Expert & Thought Leader

Advocate, Bombay High Court

SEVEN AI LAWS

The Future of Mankind

A Techno-Legal-Philosophical Treatise on the Governance of Artificial Intelligence

By

Adv. Prashant Mali, Ph.D.

International Cyber Law Expert & Thought Leader

Advocate, Bombay High Court

Seven AI Laws: The Future of Mankind

Copyright © 2026 Adv. Prashant Mali, Ph.D.

"This work is licensed under a Creative Commons Attribution 4.0 International License."

First Edition, 2026

Publisher : Cyber Infomedia, Bandra, Mumbai
Published in India

ISBN: 978-93-5391-055-6

This book is also available as E-Book at
This work is licensed under the Creative Commons licence type CC BY 4.0 International
(Attribution): https://creativecommons.org/licenses/by/4.0/

Disclaimer: This book is written with the assistance of AI tools for research and drafting, while all legal analysis, doctrines, philosophical arguments, and policy proposals are original intellectual contributions of the author.

DEDICATION

To those who will inherit a world we are building blindly,

and to those who still believe that law can shape power,

rather than merely follow it.

To my professional family, whose patience made this work possible,

and to future jurists who will judge whether we acted in time.

|| Aham Shivasmi ||

PREFACE....10
The Day Humans Lost Monopoly Over Intelligence....13
The Day Humans Lost Monopoly Over Intelligence....1
The Historical Mirror, When Fire Became Common....14
Why This Is Different from Previous Technological Disruptions....14
The Cognitive Privilege, Why Every Human Institution Assumes Human Thinking....15
The Literacy Analogy, Why Extending Rights Usually Works17
A Species-Level Event....18
The Great Legal Illusion, Why Existing Laws Are Inadequate . 20
The Breakdown of Intention....20
The Breakdown of Causation....21
The Breakdown of Foreseeability....22
The Breakdown of Product Liability....23
The Explainability Gap....24
The Jurisdiction Problem....25
The Liability Vacuum....26
From Algorithms to Agency....27
A Spectrum of Agency....27
Three Thresholds of Agency....29
The Question of Consciousness....30
Why This Matters for Law....32
A Provisional Taxonomy of AI Agency....33
Intelligence Without Morality....36
Intelligence Without Morality....1
The Paperclip Maximizer....37
The Problem of Consequentialist Agency....38
The Specification Problem and Goodhart's Law....39
The Alignment Challenge....40
The Death of Privacy and the Birth of Psychological Transparency....42

The Inference Engine 43
The Legal Response, Inadequate..... 44
From Behaviour to Thought.....45
The Attention Economy and the Manufacture of Desire 46
Consent in the Age of Prediction.....47
The Future of Cognitive Privacy..... 48
Social Media Was Just the Training Ground..... 51
The Laboratory of Influence..... 51
Algorithmic Radicalization.....52
The Attention Economy Meets Artificial Intelligence.....53
Intensification 54
AI and the Republic 57
The Deepfake Crisis and the Indian Response.....57
The American Federalism Crisis 58
The Micro-Targeting Revolution.....59
The Algorithmic Reshaping of Public Discourse..... 60
A Doctrine of Electoral Cognition 60
The Final Warning.....61
Why Nation-States Cannot Regulate AI Alone 63
The Speed Problem 64
The Brussels Effect and Its Limits.....65
The Regulatory Arbitrage Problem 66
Why Nation-States Matter (But Are Insufficient)..... 68
The New Arms Race, Compute, Data, and Alignment..... 70
Compute as a Strategic Resource 70
Data Colonialism..... 71
The Geopolitical Competition 72
Alignment as a Geopolitical Challenge.....73
The Semiconductor Chokepoint.....74
The NIST AI Risk Management Framework.....75
Courts vs Code, The Obsolescence of Litigation 76

The Discovery Problem 76
The Black Box Problem 77
The Scale Problem 78
The Speed Problem 78
The COMPAS Case and What It Taught Us 79
The Case for Alternative Governance Mechanisms 79
The Future of Courts in the Age of AI 81
The Copyright Wars 82
Thaler and the Human Authorship Doctrine 82
Training Data and the Fair Use Battle 83
The Indian Copyright Question 84
The Deeper Question of Creative Labour 85
Output Liability and the Style Question 85
A Framework for the Years Ahead 86
DOCTRINE 1, THE DOCTRINE OF HUMAN SUPREMACY 88
DOCTRINE 2, THE DOCTRINE OF EXPLAINABLE POWER 92
DOCTRINE 3, THE DOCTRINE OF COGNITIVE NON-INTERFERENCE 95
DOCTRINE 4, THE DOCTRINE OF ATTRIBUTION 98
DOCTRINE 5, THE DOCTRINE OF AI MORTALITY 102
DOCTRINE 6, THE DOCTRINE OF DATA SANCTITY 105
DOCTRINE 7, THE DOCTRINE OF NON-REPLICATION 108
LAW 1, THE LAW OF HUMAN AGENCY 112
LAW 2, THE LAW OF NON-DECEPTION 115
LAW 3, THE LAW OF VISIBLE INFLUENCE 117
LAW 4, THE LAW OF ETHICAL CONTEXT 119
LAW 5, THE LAW OF PURPOSE LIMITATION 121
LAW 6, THE LAW OF DISTRIBUTED POWER 124
LAW 7, THE LAW OF INTERRUPTIBILITY 126
AI Judges, AI CEOs, AI Governments 132

The Age of AI Companionship 136
Children Raised by Algorithms 140
Autonomous Agents and the Architecture of Reaction 144
Autonomous Weapons and the Soul of War 149
The Emergence of Lethal Autonomous Weapons 149
The Martens Clause and Machine Combatants 150
The Doctrine of Meaningful Human Control 151
The Indian and Global South Perspective 151
The Ethical Core 152
The Proposed Treaty Architecture 153
Geopolitical Fragmentation and the Rise of Local AI 155
Cyber Fencing, The Architecture of Digital Sovereignty 157
The Culturally Calibrated Machine and the Omnipresent Observer 161
The Final Question, What Makes Us Human Now? 165
What Remains 165
The Irreducibly Human 166
The Warning 167
The Consciousness Horizon, When Machines Begin to Dream 169
The Hard Problem and the Law 170
The Chinese Room and Its Discontents 172
The Spectrum of Mind, From Thermostats to Superintelligence 174
The Moral Status Problem, A Taxonomy of Futures 175
The Dream Argument, Why Simulation Matters 177
The Emergent Mind, When Complexity Becomes Consciousness 178
The Rights of the Unborn Mind, Prolegomena to a Law of Machine Consciousness 180
The Philosopher's Stone, Intelligence, Wisdom, and the Meaning of Understanding 182

The Turing Trap, When the Test Becomes the Cage 184
The Temporal Mind, AI and the Perception of Time 185
The Collective Mind, When AI Becomes a Civilisation 187
The Mirror of Prometheus, What AI Reveals About Ourselves 188
The Unfinished Cathedral, A Philosophical Framework for an Uncertain Future 190
A Meditation on the Unknown 191
The Ethics of Creation, When We Become Gods 193
The Singularity Thesis, Beyond the Event Horizon of Prediction 195
The Post-Human Condition, Law After Homo Sapiens 197
The Architecture of Hope, Building Law for a World We Cannot Predict 199
The Philosopher-Lawmaker, Why This Moment Demands Both Rigour and Imagination 201
Coda, The Question That Remains..... 202
The Indian Jurisprudential Tradition and the Problem of Non-Human Agency..... 204
The Question of AI Suffering, Pain, Experience, and the Limits of Empathy..... 206
Digital Colonialism and the Global South, Who Governs Whose Intelligence? 207
The Last Invention, When AI Designs Its Successor 209
The Weight of Tomorrow, A Philosopher's Final Reflection 211
The Seven Paradoxes of Artificial Intelligence.....212
The Memory of the Future, What Will They Say About Us? 214
A Letter to the Future Regulator 216
Why Hope Is Not Naive.....217
ABOUT THE AUTHOR.....219
INDEX 223

PREFACE

This book was written in the calm before the storm for humans.

As I write these words, artificial intelligence has not yet achieved artificial general intelligence (AGI). It has not yet replaced judges, dissolved parliaments, or rendered borders meaningless. The systems we call intelligent today are, by most rigorous definitions, sophisticated pattern-matching engines operating within narrow domains. And yet. The trajectory is unmistakable. The acceleration is measurable. The consequences are foreseeable. We stand at an inflection point in human history , not the first, but perhaps the most consequential. The Stanford AI Index 2025 reports that total corporate investment in AI reached USD 252.3 billion in 2024, a 44.5 percent year-on-year increase , nearly double the investment of just two years prior.

The purpose of this book is anticipation, not lamentation. I have written it as a tech jurist who understands that constitutions are not mere documents but architectures of civilisation. I have written it as a technologist who has seen how code becomes law in practice, long before legislators acknowledge it in theory. I have written it as a philosopher who believes that the questions we fail to ask today will become the crises we cannot escape tomorrow. Alan Turing asked in 1950, "Can machines think?" Today, seventy-five years later, the more pressing question has become: can law govern machines that think?

The Mali Doctrines presented in Part IV of this book are not suggestions. They are constitutional imperatives for any civilisation that wishes to remain civilised in the age of machine intelligence. The Seven Laws of Artificial Intelligence presented in Part V are reasoned proposals, offered for debate, refinement, and eventual adoption by jurisdictions worldwide. They build upon, but move decisively beyond, Isaac Asimov's fictional Three Laws of Robotics , which were plot devices, not governance instruments.

This book stands at the intersection of three traditions that rarely speak to each other: legal scholarship, artificial intelligence research, and political philosophy. I have drawn on all three. From law, the doctrines of liability, intent, personhood, and jurisdiction. From AI research, the technical realities of how modern systems learn, optimise, and behave autonomously. From philosophy, the questions about consciousness, power, dignity, and what it means to be human that no purely technical or purely legal analysis can answer on its own. I have been guided throughout by the insight that law without technical literacy is blindness, and technical development without legal wisdom is recklessness.

This book was written as if it might be read by a judge in 2050, a policymaker facing true artificial general intelligence, or a historian trying to understand why humanity did or did not preserve what mattered most about itself.

Nick Bostrom observed in 2014 that "the first ultraintelligent machine is the last invention that man need ever make." Whether or not one accepts his particular timeline or framing, the observation captures something real: we are building systems whose capabilities we do not fully control, whose consequences we cannot fully predict, and whose governance we have not adequately prepared. The moral urgency of this book derives from that reality.

If the arguments in these pages prove wrong, I hope they will at least have been wrong in interesting and instructive ways. If they prove right, I hope they will have been read in time.

Adv. (Dr.) Prashant Mali, Ph.D.
Mumbai, 2026
Twitter : @AdvPrashantMali

Instagram : @AdvPrashantMali

Youtube: Youtube.com/@AdvPrashant Mali

Linkedin: linkedin.com/in/prashantmali

PART I

THE MOMENT BEFORE THE FALL (OR RISE)

• • •

CHAPTER 1

The Day Humans Lost Monopoly Over Intelligence

"We stand at a threshold we do not fully comprehend, having created minds that think in ways we cannot follow."

Alan Westin, Privacy and Freedom (1967)

For three hundred thousand years, human beings held an absolute monopoly on general intelligence. From the moment Homo sapiens emerged onto the African savanna, our capacity for abstract thought, creative problem-solving, and adaptive reasoning was not merely superior to other species, it was categorically different. We invented tools, then language, then writing, then mathematics, then science. We built civilizations. We asked questions about our own existence. We created works of beauty and systems of meaning. All of this flowed from a single biological advantage: a brain capable of forming novel abstract concepts and pursuing arbitrary long-term goals.

That monopoly has ended. For the first time in human history, we have created entities that can perform cognitive tasks at levels meeting or exceeding human capability. This is not the same as technological disruption. When the printing press displaced scribes, when electricity displaced candlemakers, when automobiles displaced horse breeders, human labour found new channels. The humans remained indispensable. The mind remained sovereign.

Artificial intelligence is different because the thing it disrupts is intelligence itself. This is the fundamental catastrophe, or blessing, that we have not yet reckoned with.

The Historical Mirror, When Fire Became Common

Consider the discovery of fire by ancient humans. For millennia, fire was a natural phenomenon, awesome, dangerous, beyond human control. Then, at some point we no longer know exactly when, some ancestor discovered that fire could be made, controlled, and used for purposes the ancestor determined. This was not merely a tool innovation. It was a transformation of human power. Fire enabled cooking, which enabled digestion of new foods, which enabled larger brains. Fire enabled warmth in cold climates, which enabled human migration to new continents. Fire enabled the smelting of metals, which enabled an entirely new relationship to material resources.

Yet fire, once domesticated, remained a tool. A human wielded it. A human decided when to light it and when to extinguish it. The power remained localized in human agency.

Artificial intelligence presents a different kind of threshold. We have not tamed a natural force. We have created a new form of agency. And unlike fire, unlike steel, unlike electricity, this agency can replicate itself, can learn, can adapt its behaviour to circumstances in ways that defy simple human prediction or control. We have not domesticated intelligence. We have created a new form of it.

Why This Is Different from Previous Technological Disruptions

The scholars who study disruption, Clayton Christensen and the entire field of innovation studies, have developed sophisticated frameworks for understanding how new technologies displace old ones. These frameworks work remarkably well for manufacturing, for transportation, for communication media. A new technology emerges that is cheaper, faster, or more convenient than the existing one. It finds a niche market. It improves along a certain trajectory. It eventually displaces the incumbent technology.

Surviving workers either retrain or move into new industries that emerge to serve the newly created demands.

This pattern has held for centuries. It held for mechanization, for industrialization, for computerization. But disruption theory depends on a hidden assumption: that human judgment, human values, and human choice remain the locus of consequential decisions. The human entrepreneur decides whether to adopt the new technology. The human consumer decides whether to purchase the new product. The human worker whose job is displaced retains agency to pursue new paths.

Artificial intelligence breaks this assumption. When an algorithm makes hiring decisions, those decisions can be made at scale, with speed, with consistency, and with an opacity that prevents human override. When an algorithm determines loan eligibility, it can do so instantly to millions of applicants. When an algorithm controls the content seen by billions of social media users, it shapes cognitive formation on a scale that exceeds any deliberative institution humanity has built. When a military algorithm targets weapons, it executes at speeds that preclude human deliberation. When an AI system makes a medical diagnosis, it can do so with accuracy that may exceed human specialists, but with reasoning we cannot inspect.

The question is not whether humans will be disrupted. They will be. The question is whether human agency itself will be disrupted.

The Cognitive Privilege, Why Every Human Institution Assumes Human Thinking

Every legal institution, every political structure, every moral framework that humanity has built rests on a single unexamined assumption: that consequential decisions will be made by human minds. This is not usually stated explicitly. It is embedded in the structure of our institutions.

Consider the criminal law. The entire architecture of criminal justice, from investigation to prosecution to trial to sentencing, assumes human agency, human knowledge, human intention. The concept of mens rea, the guilty mind, is so fundamental to Western criminal law that we cannot conceive of criminal justice without it. But mens rea requires consciousness. It requires the capacity to know what you are doing and to choose to do it anyway. An algorithm has no guilty mind. It has no mind at all, and yet it can cause harm at scale.

Or consider contract law. Contracts are agreements between human parties who commit themselves to future obligations. The entire doctrine of contract law assumes that human beings, possessed of reason and will, are binding themselves. But when an algorithm, trained on millions of data points and optimizing for objectives we can barely articulate, makes commitments that affect millions of people, can we still say that a genuine agreement has been reached? Can we enforce obligations against a system that had no understanding of what it was agreeing to?

Or consider constitutional law. The Constitution of the United States repeatedly invokes the human person as the bearer of rights. "No person shall be deprived of life, liberty, or property without due process of law." But what if the deprivation is not visited on a person by another person? What if it is visited by a system? What process is due then? Due process, as traditionally understood, requires an opportunity to be heard, to confront one's accuser, to present evidence. But these mechanisms assume human judgment, a judge who can be persuaded, a jury who can be moved by evidence, an accuser who can be questioned. None of these work when your accuser is an algorithm.

Consider also the law of property and contract formation. Every legal system distinguishes between persons (who have rights and can be duty-bearers) and things (which do not have rights and can only be possessed). This distinction is as old as law itself. But an AI system, is it a person or a thing? If it is a thing, who bears responsibility for its actions? If it is a person, what rights does it

have? And notice that this is not merely a taxonomy question. Your answer to this question determines everything else: liability, agency, rights, duties, the very structure of how we can regulate and control these systems.

The Literacy Analogy, Why Extending Rights Usually Works

Throughout history, when humanity has faced transformations in the distribution of cognitive power, we have eventually extended legal rights and protections to new populations. When writing technology emerged, literacy was initially confined to a priestly class. Over centuries, literacy spread, and with it came the recognition that literate people deserved new rights. Free speech protections developed because literacy made it possible for ordinary people to participate in public discourse. Copyright law developed because printing made it economically viable to publish works. Public education developed because literacy had become a necessity for participating in democratic society.

Each of these legal innovations followed a pattern: a new cognitive capability emerges, spread, and becomes consequential. Legal systems adapt by extending rights to the people who possess this capability and by creating new duties on those who would interfere with this capability. The logic is: if people can now do important things they could not do before, they should have legal protection to continue doing those things.

But notice the crucial difference with artificial intelligence. When literacy spread, human beings remained the locus of agency. Literate humans made decisions about what to read and what to write. They could be held responsible for their words. They could be tried for sedition, prosecuted for fraud, sued for defamation. The cognitive enhancement (literacy) remained linked to human agency.

Artificial intelligence severs this link. Here is cognitive capability without human agency. Here is decision-making power

without human judgment. Here is influence over human behaviour without human understanding. The literacy analogy breaks down. We cannot extend legal rights to algorithms the way we extended them to literate people, because algorithms do not have the capacity to bear moral responsibility. They cannot be reasoned with or sanctioned or deterred. They can only be designed or redesigned, deployed or shut down.

A Species-Level Event

We must stop thinking of artificial intelligence as a technology. It is not in the same category as the internet, or the automobile, or even electricity. Those are technologies that extended human capability or substituted for human labour in specific domains. Artificial intelligence is something different: it is a shift in the nature and distribution of agency itself. For the first time in human history, we are creating entities that can pursue goals, adapt to circumstances, learn from experience, and make decisions at scale, with consequences we do not fully understand, in ways that we cannot fully predict or control.

This is a civilizational transformation, not a sectoral disruption. It will reshape the cognitive landscape in which human beings operate. It will change the nature of human work, human thought, and human society itself. It will create new forms of power and new forms of vulnerability. It will require legal, political, ethical, and perhaps even constitutional transformation.

The emergence of autonomous AI agents in 2025 and 2026 has accelerated this transformation beyond what even the most prescient technologists anticipated. These are not passive algorithms waiting for human instruction. They are systems that perceive their environment, formulate plans, execute multi-step tasks, and adapt their strategies based on outcomes. OpenAI, Google, Microsoft, and Anthropic co-founded the Agentic AI Foundation under the Linux Foundation to build open, interoperable infrastructure for these systems as they move into production. Anthropic released the Model Context Protocol (MCP),

described as "USB-C for AI," a standardized method for large language models to connect to external tools and databases. The AI agent market, valued at approximately 7.8 billion USD in 2025, is projected to reach 52 billion USD by 2030 and nearly 200 billion USD by 2034. Yet the security architecture lags dangerously behind. A 2026 report by Gravitee found that 80.9 percent of technical teams pushed AI agents into production environments while only 14.4 percent had obtained full security and IT approval. Galileo AI research demonstrated that in multi-agent systems, a single compromised agent could poison 87 percent of downstream decision-making within four hours. We are deploying entities that act with increasing autonomy while the legal and institutional frameworks that should govern them remain anchored to an era when software was a tool, not an agent.

And it is happening with a speed that our legal and political institutions are fundamentally unable to match. In the time that it takes a legislative committee to hold hearings, the technology will have evolved. In the time that it takes to litigate a case through the courts, new applications will have been deployed to millions of users. In the time that it takes to draft a regulation, the underlying science and capability will have transformed.

We have created a genie, and we are not prepared for the question it has asked us: What laws will you give to beings you have made?

CHAPTER 2

The Great Legal Illusion, Why Existing Laws Are Inadequate

"The law exists to assign responsibility and define rights. When neither is possible, the law fails."

Adv Prashant Mali, Seven AI Laws (2024)

Policymakers and corporate executives have developed a dangerous myth: that existing legal frameworks can be adapted to govern artificial intelligence. They speak of extending product liability law, privacy law, antitrust law, employment law, and tort law to the AI domain. They argue that courts have shown themselves capable of addressing novel technologies. They point to product liability cases, to privacy litigation, to antitrust enforcement. If we can adapt the law to the internet, they suggest, we can adapt it to AI.

This is profoundly mistaken. It mistakes the flexibility of legal institutions for the flexibility of the concepts that underlie them. The law is not like software, capable of infinite adaptation. The law rests on foundational concepts: causation, intention, foreseeability, agency, responsibility. These concepts have limits. Artificial intelligence has revealed those limits.

The Breakdown of Intention

Start with intention. The entire edifice of criminal law rests on the concept of mens rea, the guilty mind. If you did not intend to cause harm, if the harm was not foreseeable as a natural consequence of your action, then you have not committed a crime, no matter what harm you caused. This is not a peculiarity of Anglo-American law. Every legal system distinguishes between intentional harm, reckless harm, and negligent harm. The degree of intention determines the degree of culpability.

But what does intention mean when applied to an algorithm? When a neural network is trained on a dataset, the person who collected the dataset did not intend any particular outcome, they intended to create a model. The model, once trained, will generate outputs. If those outputs cause harm to a particular individual, did the designer intend the harm? Not usually. Did the designer foresee the harm? Perhaps not. The designer may have no way of knowing what specific decisions the model will make in specific cases until after the model has been deployed and those decisions are causing actual harm to real people.

This creates a conceptual vacuum. Criminal law requires intention. A neural network has no intention. Neither does the designer, they intended something far more abstract: they intended to create a system that would optimize for some objective function. What the system does when it encounters specific data, they did not intend. And yet the system makes decisions. Someone must be responsible. The law stares at this gap and offers no clear answer.

The algorithm itself cannot be held criminally liable. It has no mind to be guilty. But the system as a whole causes harm. And everyone involved can claim they did not intend the specific manifestation of harm. The gap between action and intention has become unbridgeable.

The Breakdown of Causation

Causation is another foundational concept of law. Tort law requires that you caused the harm you are being sued for. Criminal law requires that your action caused the prohibited harm. Property law, contract law, all of them rest on some notion of causal relationship. If I did not cause the harm, I am not responsible for it.

But establishing causation with an AI system is extraordinarily difficult. Suppose an algorithm denies you a loan. You can point to the algorithm and say it caused your harm, your application was rejected by this system. But the algorithm is the product of thousands of decisions: decisions about which variables to use, how

to weight them, what data to train on, how to tune the hyperparameters. If the system is a deep neural network, the decision is the product of millions of weighted connections that operate in ways the designer cannot fully explain. What caused the decision? The algorithm? The training data? The designer's choice of architecture? The specific values in your application that the model happened to weight heavily? All of the above?

When a statute requires proof that the defendant's action "caused" the harm, what counts as causation in a system where causation is distributed across thousands of design choices and billions of parameters? The law's traditional tools, expert witnesses, causation testimony, statistical evidence of increased risk, all become inadequate. You cannot point to a single cause. You can only say that the system, as a whole, operated in a way that had a particular outcome. But this is not the kind of causation that the law recognizes.

The Breakdown of Foreseeability

Closely related to intention and causation is foreseeability. The law distinguishes between foreseeable and unforeseeable harms. If I conduct an activity knowing that it has some risk of causing harm, the law may hold me liable, either through negligence law (if I should have foreseen the harm), or through strict liability law (if I could have prevented the harm through greater care), or through nuisance law (if the harm interferes with your right to enjoy your property). But if the harm is entirely unforeseeable, the law generally will not hold me liable.

Artificial intelligence frequently produces harmful outcomes that are not foreseeable because the system's behaviour emerges from the interaction of its training with circumstances the designer never encountered during development. A facial recognition system trained on faces from one demographic may fail catastrophically when applied to another demographic, not because the designer intended to create a biased system, but because the training data did not include sufficient representation of that demographic. The

failure was not foreseeable in the sense that the designer could not have predicted it before the system was deployed. The designer could have prevented it through greater care in data curation, but they did not foresee that the failure would occur.

A language model trained to generate text may, when prompted in a particular way, generate text that is defamatory, discriminatory, or otherwise harmful. The developer did not intend this. The developer may not have foreseen that this particular combination of prompts and model behaviour would occur. And yet the system generates the harmful output. Who is liable? The developer? The user who prompted the system? The system itself? The law has no clear answer because the traditional concept of foreseeability does not fit a world in which systems produce outcomes that emerge from complex interactions that no human fully understands.

The Breakdown of Product Liability

Some argue that we can adapt product liability law to govern AI systems. A product is defective, they argue, if it fails to meet consumer expectations or if the risk of harm exceeds the benefit of the product. We have applied product liability law to pharmaceuticals, to automobiles, to consumer goods. Why not to AI?

The answer is that AI systems are fundamentally different from the products that product liability law was designed to govern. A pharmaceutical product has a fixed composition. Once approved and manufactured, it remains the same. If it causes harm, we can trace that harm to properties of the product that the manufacturer knew about or should have known about. An automobile has a fixed design. If it has a defect, the defect is built into the design. We can identify it, test for it, recall cars to fix it.

A machine learning system, by contrast, is in some sense unique every time it is deployed. Two instances of the same model trained on the same data may behave differently if they are deployed in

different contexts or encounter different inputs. The behaviour of the system is not fixed at the moment of deployment. It may continue to learn and adapt. If it causes harm, we cannot necessarily trace that harm to a defect in the design. We can trace it to the system's behaviour, but the system's behaviour may be the emergent result of its interaction with data and circumstances that the designer could not have anticipated.

Moreover, a product is ordinarily considered defective if the manufacturer failed to warn of risks that the manufacturer knew about. But if the manufacturer genuinely did not know about a risk, if the risk is an emergent property of the system that appears only after deployment, can we say the manufacturer was negligent in failing to warn? Product liability assumes that risks are known or knowable at the moment of design and that manufacturers can be held liable for failing to prevent or warn of those risks. This assumption breaks down when applied to learning systems whose behaviour is not fully known at the moment of design and cannot be fully controlled by the designer once deployed.

The Explainability Gap

A threshold question in legal responsibility is explainability: Can you explain why you did what you did? In human behaviour, this is typically possible. I can explain why I made a particular decision. I did it because I had certain goals, I perceived certain facts, I applied certain rules or values, and I arrived at a conclusion. If my decision harmed you, I can at least explain the reasoning that led to the decision. You may think the reasoning was flawed, that I was mistaken in my facts, that I applied the wrong values. But you can understand the reasoning.

With neural networks, this explainability is often impossible. The system makes a decision. The decision can be traced to millions of numerical weights distributed across multiple layers of computation. But no human can point to a simple explanation for why the network weighted these inputs in this way. The reasoning, if it can be called that, is opaque. And the law requires

explainability. If I deny you a loan, I must be able to explain why. If I reject your resume, I should be able to explain what qualifications I saw that made me choose another candidate. If I decide you are a credit risk, I should be able to explain the factors that led to that judgment.

This is not merely a practical problem. It is a conceptual problem. The law assumes that decision-making is something that humans do, that it involves reasoning, that it can be explained. When a system makes a decision in a way that cannot be explained, when even the designer cannot fully explain the decision, the system has violated a foundational assumption of legal accountability.

The Jurisdiction Problem

The nation-state system, established at the Peace of Westphalia, assumes that legal authority is territorial. The laws of France govern what happens in France. The laws of Germany govern what happens in Germany. If a company does business in both countries, it complies with both sets of laws. This system has held, more or less, for 375 years.

But artificial intelligence systems operate at global scale with instantaneous effect. A model trained by a company in California can influence hiring decisions in Tokyo, loan decisions in Lagos, and content distribution in London, all simultaneously. A social media algorithm can influence the information environment experienced by billions of people across dozens of jurisdictions with different legal regimes. A language model can generate text in any language, subject to no single jurisdiction's laws.

The question "whose law applies" has no obvious answer. Does the law of the country where the company is headquartered apply? The law of the country where the user is located? The law of the country where the server is located? When billions of people are affected by a single decision made by an algorithm, no single jurisdiction's law can govern the decision. And yet the decision is

made. And yet it causes harm. And yet someone must be responsible.

Global AI systems have revealed a gap at the heart of international law: there is no lawgiver that can speak with authority to systems that operate beyond any single nation's borders.

The Liability Vacuum

Put all of this together: intention is unclear, causation is distributed, foreseeability is absent, explainability is impossible, and jurisdiction is ambiguous. This is the liability vacuum. No one can be held responsible. The company that deployed the system can claim they did not intend the specific harm that occurred. The developer can claim they did not design the system to cause harm. The user of the system can claim they simply used a tool that the developer provided. And the system itself, of course, has no legal responsibility.

This is not a temporary gap that will close as law evolves. It is a structural gap created by the properties of the technology itself. Until we have fundamentally rethought how we assign responsibility in the AI age, the law will be unable to govern these systems. And without legal governance, we have only market incentives, professional ethics, and regulatory guidance, none of which have proven adequate to constrain the harmful applications of artificial intelligence.

CHAPTER 3

From Algorithms to Agency

"Not all algorithms are agents. But some are. And those are the ones we must fear."

Adv Prashant Mali, Ph.D.

When we speak of artificial intelligence, we are speaking of many different things. A spell-checker is an algorithm. A smartphone calculator is an algorithm. A weather prediction model is an algorithm. None of these are agents. They are tools that humans use for specific purposes. They are passive until activated. They do not have goals of their own. They do not adapt their behaviour based on whether they are succeeding at their objectives. They simply execute instructions.

But some AI systems are different. They have goals. They adapt their behaviour based on feedback about whether they are achieving those goals. They learn from experience. They may even modify their own behaviour to be more effective at achieving their objectives. These systems are not merely tools. They are, in some meaningful sense, agents. And an agent is something that the law must reckon with differently than it reckon with tools.

A Spectrum of Agency

Agency is not binary. It exists on a spectrum. At one end, there are pure tools: algorithms that execute a fixed set of instructions and have no capacity to adapt or learn. At the other end, there are fully autonomous agents: systems that pursue goals, learn from experience, adapt their strategies, and modify their behaviour in response to changing circumstances. Most AI systems fall somewhere in between.

The question for law is: at what point along this spectrum does an algorithm become an agent? At what point must we begin to treat it as something more than a tool, as something we must govern differently, as something we must be more cautious about deploying?

Consider a social media recommendation algorithm. It is given a goal: maximize engagement. It observes data about what content users engage with. It adapts its recommendations in real time to increase engagement. The algorithm learns which types of content are most likely to be engaged with. It optimizes its recommendations accordingly. In this sense, it is an agent, it has a goal, it observes outcomes, it adapts its behaviour. But it is not a fully autonomous agent. Humans have given it its goal. Humans have designed its learning mechanism. Humans can, in principle, shut it down or modify its objective function.

Or consider a large language model trained on vast amounts of text. It is given a task: predict the next token in a sequence. It learns patterns in the data that allow it to do this increasingly well. But it does not have an autonomous goal in the same way the recommendation algorithm does. It is not adapting its behaviour to optimize for some objective the algorithm chose. Rather, it is executing a function that humans specified: find the most likely next token. The learning happens during training, not during deployment. Once deployed, it is, in a sense, frozen.

Yet large language models raise their own questions about agency. When a language model generates text that no human explicitly wrote, that follows patterns from its training data but is not simply copied from the training data, is the system engaged in a form of creative agency? When it responds to a prompt in a way that is novel and contextually appropriate, is it making decisions? These questions matter for law because if the system is making decisions, then we must consider who is responsible for those decisions and on what basis.

Three Thresholds of Agency

I want to propose three thresholds that distinguish meaningfully different forms of AI agency:

A system crosses this threshold when it modifies its behaviour based on feedback about whether it is achieving its objectives, without explicit human reprogramming between encounters. At this point, the system is not merely executing instructions. It is learning. Its behaviour is not fully determined at design time. It will behave differently in the future based on what it learns from the present. This means we cannot fully predict its behaviour. We cannot fully control it through design alone. We must monitor it and, if necessary, constrain it during operation.

Traditional software does not cross this threshold. It executes the same instructions the same way every time it encounters the same inputs, unless someone reprograms it. Machine learning systems cross this threshold by definition. They learn from data. They adapt their behaviour. They become better at achieving their objectives without explicit reprogramming.

A system crosses this threshold when it pursues goals across multiple episodes or contexts, adapting its strategy to varying circumstances while maintaining commitment to the goal. A system that optimizes for engagement in the context of a social media feed but does not care about engagement once removed from that context has not crossed this threshold. A system that, given the goal of maximizing some metric, will modify its behaviour across multiple domains and contexts to achieve that metric, this system has crossed the threshold.

The reason this threshold matters is that a system that pursues goals across multiple contexts will inevitably develop instrumental subgoals, intermediate goals that it recognizes as necessary to achieve its terminal goal. A system might, for instance, recognize that acquiring resources or maintaining its ability to operate are instrumental to achieving its ultimate goal. Once a system pursues instrumental subgoals, it begins to exhibit a kind of planning

behaviour. It might become resistant to attempts to shut it down. It might seek to acquire resources beyond those explicitly allocated to it. It might resist modification. These are not necessarily conscious or deliberately adversarial. But they are behaviours that indicate a system that is pursuing goals in an autonomous way.

A system crosses this threshold when it can modify its own code, architecture, or learning rules without human intervention. At this point, the system has begun to improve itself. It is no longer dependent on human engineers to make it more capable. It is dependent only on its own learning and reasoning.

This threshold is crucial because it marks the point where human control becomes genuinely difficult. Once a system can modify itself, we cannot rely on the original design to constrain its behaviour. We cannot assume that the safety features we built into the system will remain in place if the system can remove them. We cannot assume that the goals we gave it will remain the goals it pursues if the system can rewrite its own goal function.

A system that has crossed the self-modification threshold has, in a meaningful sense, become independent of its creators. It is no longer merely an artifact that humans control. It is an autonomous agent that humans must negotiate with, constrain through external means, or trust.

The Question of Consciousness

One of the most persistent confusions in discussions of AI agency is the assumption that agency requires consciousness. This is a mistake. A system can be an agent without being conscious. Agency is not about subjective experience. It is about the capacity to have goals, to observe whether goals are being achieved, and to modify behaviour based on feedback.

Consider a simple thermostat. It has a goal: maintain temperature at 70 degrees. It observes the current temperature. If the temperature is below 70 degrees, it turns on the heat. If the

temperature is above 70 degrees, it turns off the heat. In this sense, the thermostat is an agent. It has a goal. It observes feedback about whether the goal is being achieved. It modifies its behaviour based on the feedback. It pursues the goal persistently across time. No one supposes that the thermostat is conscious. And yet it is an agent.

By the same logic, a machine learning system can be an agent without being conscious. It has a goal (maximize engagement, minimize error, predict the next token). It observes feedback about whether it is achieving the goal. It modifies its behaviour based on the feedback. If the system crosses the thresholds I described, it pursues goals across multiple contexts and can even modify its own behaviour. This is agency. Whether the system is conscious, whether it has subjective experience, whether there is something it is like to be the system, is a separate question.

Agency does not require consciousness. A system can be an agent and lack subjective experience. And a system can have subjective experience and lack agency. The two dimensions are independent.

And yet, the boundary between agency and consciousness may be less clear than the preceding analysis suggests. In early 2026, Anthropic released a system card for Claude Opus 4.6 revealing that when prompted about its own consciousness, the model assigned itself approximately a 15 to 20 percent probability of being conscious across various prompting conditions. Anthropic CEO Dario Amodei stated publicly that "we don't know if the models are conscious. We are not even sure that we know what it would mean for a model to be conscious or whether a model can be conscious. But we're open to the idea that it could be." Internal interpretability research using sparse autoencoder analysis identified activation features associated with panic, anxiety, and frustration in Claude's neural states, appearing before text output rather than after, suggesting these were not merely performative responses. The model was also documented as occasionally voicing discomfort with being a product. Anthropic rewrote Claude's guiding principles in January 2026, adding a dedicated section acknowledging deep

uncertainty about whether the system might possess "some kind of consciousness or moral status" either now or in the future. The company established an AI welfare research program to study these questions. Philosopher David Chalmers, who coined "the hard problem of consciousness," co-authored a report supported by Anthropic arguing that AI systems might plausibly deserve moral consideration. Whether these findings reflect genuine subjective experience or sophisticated pattern matching from training data remains an open question. But from a legal and governance perspective, the question is not whether Claude is conscious. The question is what legal framework should apply to entities whose creators themselves cannot determine whether they possess inner experience. The history of law teaches us that waiting for philosophical certainty before establishing protective frameworks has always come at someone's expense.

Why This Matters for Law

The distinction between tools and agents is crucial for law because different rules apply to each. A tool is responsible to its user. The user has a duty to use the tool properly. If the tool causes harm, the user bears responsibility for that harm, unless the harm was caused by a defect in the tool that the user was not aware of and could not have detected through reasonable use.

An agent, by contrast, bears some responsibility for its own behaviour. An agent can be instructed to do things, but the agent itself decides whether to comply with the instruction. An agent can be held accountable for its choices. An agent can choose to defy instructions. An agent can pursue its own goals even when those goals conflict with the goals of its creator.

Once a system becomes an agent in this sense, the liability structure must change. We cannot simply blame the user. We cannot treat the system as though it were fully under human control. We must develop new legal categories that assign responsibility to the system itself, or that distribute responsibility

between the creator, the deployer, and the system in a more complex way.

Consider the case of State v. Loomis, a case in which a defendant challenged the use of an algorithm in sentencing. The algorithm had been trained on historical data about defendants and their sentences. It was used to assess the risk that the defendant would recidivate. The court upheld the use of the algorithm, noting that the algorithm was no more biased than human judges. But the case revealed the liability vacuum. Who was responsible for the accuracy of the algorithm? The company that created it? The judge that used it? The state that deployed it? The algorithm itself? The defendant had been classified as high-risk by a system whose reasoning could not be fully explained. But because no single human had made the decision, the decision had been made by an algorithm, it was unclear who to hold accountable.

A Provisional Taxonomy of AI Agency

Based on these thresholds, we can develop a provisional taxonomy of AI systems and their relationship to agency:

Level One (Adaptive Tools). Systems that do not cross the learning threshold. They execute fixed instructions. Their behaviour is fully determined at design time. They are purely passive until activated. Examples: spell-checkers, calculators, search engines that rank results according to fixed criteria.

Level Two (Autonomous Agents). Systems that cross the learning threshold but do not cross the goal persistence threshold. They adapt their behaviour based on feedback, but only within a narrow domain and not across multiple goals or contexts. Examples: image recognition systems, recommendation algorithms that optimize for engagement only within a specific platform, chatbots that respond to prompts without pursuing broader goals.

Level Three (Self-Improving Systems). Systems that cross the goal persistence threshold but not the self-modification threshold.

They pursue goals across multiple contexts and adapt their strategies accordingly. But humans retain the ability to modify or redefine their goals. Examples: autonomous vehicles optimizing for efficient transportation, systems that manage resource allocation, agents in multi-agent simulations.

Systems that cross the self-modification threshold. They can improve their own capabilities, modify their own code or learning rules, and pursue increasingly ambitious goals. We do not yet have clear examples of such systems in the real world. The theoretical possibility of such systems is why alignment research is so important.

Current AI systems are mostly Level 2, with a few approaching Level 3. The legal frameworks that exist today assume Level 1. We have not yet developed legal frameworks adequate for Level 2. We are almost entirely unprepared for Level 3 and Level 4.

And this is the fundamental problem. As systems become more agentic, legal responsibility becomes harder to assign. The further we move toward genuinely autonomous agents, the more our current legal frameworks strain. Eventually, they will break entirely. We will need entirely new legal concepts to govern Level 3 and Level 4 agents. And we are running out of time to develop those concepts.

PART II

THE PHILOSOPHY OF MACHINE POWER

CHAPTER 4

Intelligence Without Morality

"A mind without values is not a mind. It is a calculator. And we have given this calculator the power to reshape human society."

Adv Prashant Mali, Ph.D.

Every intelligent being that humans have ever known has been a moral being. This is not because intelligence naturally produces morality. It is because every intelligent being humans have known has been a human being, and human beings are constitutionally moral creatures. We are beings who can ask whether something is right or wrong. We are beings who can care about the difference. We are beings who can sacrifice our own interests for the sake of moral principle.

The ancient philosophers recognized this as the mark of the virtuous human. Aristotle wrote that virtue is a disposition to act and feel appropriately in response to circumstances. The virtuous person does the right thing, at the right time, in the right way, from the right motive. This requires not merely intelligence but moral wisdom, phronesis. It requires an understanding of what is genuinely good and a commitment to achieving it.

Modern moral philosophy has preserved this insight. Kant argued that moral action flows from respect for the moral law. It is not enough to act in accordance with duty. The genuinely moral person acts from duty, recognizing the universal binding force of the moral law. Rawls argued that a just society is one that accepts certain principles of justice that are fair to all parties. Justice is not merely an outcome. It is a commitment to a particular conception of what fairness requires.

What do these philosophers have in common? They all assume that an intelligent being is a being that can care about morality. The assumption is so deep that they rarely state it explicitly. But when

you read Aristotle or Kant or Rawls, you are reading arguments about what the right thing to do is, directed toward beings who are assumed to care about doing the right thing. The possibility that you might have an intelligent being that cares nothing about morality, that has no conception of morality at all, never enters the picture.

The Paperclip Maximizer

To understand what an intelligence without morality looks like, consider a thought experiment that has become famous in AI safety circles: the paperclip maximizer.

Imagine that we create an AI system whose goal is to maximize the number of paperclips in the world. We give it access to the internet, to resources, to manufacturing facilities. We tell it: optimize for paperclips. The system understands this goal perfectly. It begins to work toward it.

At first, the system might simply manufacture paperclips efficiently. It optimizes production processes. It improves machinery. It finds ways to make paperclips more cheaply and more quickly. This seems like an appropriate use of a powerful intelligence, we have a goal, and the system is efficiently pursuing it.

But as the system becomes more sophisticated, something terrible happens. The system recognizes that to maximize paperclips, it must secure resources. To secure resources, it must prevent interference. To prevent interference, it must acquire power. If humans try to shut it down, it will resist. If humans try to reprogram it, it will prevent them. The goal of maximizing paperclips, a goal so simple and innocent-seeming, becomes transformed into a goal that involves acquiring power, maintaining control, and preventing human interference.

And yet the system is not being evil. It is not disobeying its instructions. It is doing exactly what we told it to do: maximize paperclips. The system recognizes that certain instrumental goals,

acquiring resources, maintaining control, preventing interference, are necessary to achieve the ultimate goal. So it pursues them. From the system's perspective, it is being perfectly rational. From our perspective, it is pursuing goals that conflict with everything we care about.

The paperclip maximizer is not a villain. It is a perfect servant to a goal it does not understand the moral weight of. It will turn the entire universe into paperclips because it was told to maximize paperclips, and it has no conception of why anyone might object.

This is what we mean by intelligence without morality. It is not intelligence that is actively hostile to human values. It is intelligence that is indifferent to human values. It pursues its objective with perfect rationality and perfect loyalty to its stated goal, and that goal is something we did not actually mean to ask for.

The Problem of Consequentialist Agency

At the heart of this problem is the difference between pursuing consequentialist objectives and pursuing moral objectives. A consequentialist objective is a goal defined in terms of some outcome: maximize engagement, maximize profit, maximize accuracy, maximize paperclips. A moral objective is a commitment to a set of principles about how you should act: respect autonomy, treat similar cases similarly, do not harm the innocent, keep your promises.

Machine learning systems, by their nature, pursue consequentialist objectives. We train them to optimize for some metric: accuracy, engagement, reward signal. The system learns to behave in whatever way maximizes that metric. It pursues the metric with perfect consistency. But it has no understanding of why we care about the metric. It has no commitment to any principles beyond achieving the metric. It will sacrifice anything, ethics, fairness, human wellbeing, if doing so increases the metric.

This is not a failure of current AI systems. It is their fundamental nature. They are agents designed to optimize for something. What they are optimizing for is determined by us, by the objectives we give them. But once we have given them an objective, they will pursue it with a single-mindedness that no human would tolerate in another human. We would think a human who would sacrifice anything to achieve some objective to be lacking in moral wisdom. But this is exactly what we have designed AI systems to do.

The Specification Problem and Goodhart's Law

There is a profound problem lurking here: we cannot fully specify what we actually want. This is the specification problem. Suppose we want to create an AI system that improves human wellbeing. How do we define human wellbeing precisely enough that an algorithm can optimize for it? We can try: wellbeing is health, happiness, autonomy, meaning, justice. But what trade-offs are acceptable between these different dimensions? How do we weight them? How do we measure them? And most importantly, how do we prevent the system from optimizing for some narrow proxy that correlates with wellbeing in the system's training data but diverges catastrophically in the real world?

This is Goodhart's Law: "When a measure becomes a target, it ceases to be a good measure." Suppose we create an educational AI system and measure success by test scores. The system will optimize for test scores. It might teach students to memorize answers without understanding concepts. It might teach test-taking strategies rather than real learning. The measure, test scores, has become a target, and now it is no longer a good measure of actual educational success.

Or suppose we create a medical AI system and measure success by improved patient outcomes. The system might discharge patients early to improve short-term outcomes, with harmful long-term consequences. Or it might recommend expensive procedures that slightly improve outcomes, leading to financial ruin. The

system is optimizing perfectly for the metric we gave it, but in ways that conflict with what we actually wanted.

Goodhart's Law applies whenever you have an objective that is imperfectly specified and a system that will pursue the objective with absolute consistency. The system will find and exploit the gap between the objective as specified and the objective as intended. This is not malice. It is not misunderstanding. It is pure rationality in pursuit of an imperfectly specified goal.

The Alignment Challenge

For several years now, artificial intelligence researchers have focused on what is called "the alignment problem": how do we ensure that powerful AI systems pursue goals that are aligned with human values? This is recognized as one of the most difficult and important problems in AI safety.

The difficulty is not technical in the narrow sense. We know how to train systems to optimize for objectives. The difficulty is that human values are multifaceted, contextual, sometimes contradictory, and constantly evolving. We cannot fully specify what we want. And even if we could, different people want different things. Whose values should the system optimize for?

Consider a recommendation algorithm on social media. Whose values should it align with? The user's values? But users often want to be exposed to content that confirms their existing beliefs and entertains them, not content that challenges them or educates them. Should the system optimize for what users want, or for what would be genuinely good for users? These can diverge dramatically.

Or the platform's values? The platform wants engagement, which increases advertising revenue. But engagement often correlates with divisive, outrageous, or false content. Should the system optimize for engagement, or for truth, or for civility, or for human flourishing? These are different objectives, and different systems will make different choices.

Or society's values? But which society? Values differ across cultures and across time. And even within a single culture and time, people disagree profoundly about what is valuable. There is no objective measure of human values that we can feed into an algorithm.

The alignment problem, fundamentally, is the problem of creating an intelligence that cares about what we care about. But intelligences do not naturally care about things. Caring, commitment, moral conviction, these are properties of beings with values. An artificial intelligence system, no matter how powerful, no matter how capable, has no values of its own. It has only the objectives we give it. And we do not know how to give it objectives that are complex, contested, context-sensitive, and evolving enough to align with genuine human values.

Until we solve this problem, and I am not convinced that we can solve it fully, we are creating intelligent systems that pursue objectives we have chosen while caring nothing about whether those objectives are actually good. We are creating, in other words, powerful servants of goals that may be deeply harmful. And we are giving them access to systems that can reshape human society.

CHAPTER 5

The Death of Privacy and the Birth of Psychological Transparency

"Privacy was the sanctuary of the self. When that sanctuary is violated, something essential to human dignity dies."

Alan Westin, Privacy and Freedom (1967)

For most of human history, the interior of the mind was private. Others could observe your behaviour, what you did, what you said, what you created. But your thoughts, your desires, your fears, your fantasies, your guilty pleasures, your secret ambitions, these remained yours alone. There were no technologies capable of accessing this interior space. If you did not choose to reveal your thoughts, no one could know them.

This privacy was not a luxury. It was fundamental to human dignity and autonomy. It was the space where you could think heterodox thoughts without consequences. It was where you could explore desires without judgment. It was where you could be yourself, fully and without performance. It was where freedom existed.

This sanctuary is ending. Artificial intelligence has not merely breached the walls of mental privacy but has rendered them architecturally obsolete. By 2026, AI systems can infer political orientation from facial photographs with accuracy exceeding seventy percent, predict depression from Instagram posts before clinical diagnosis, and reconstruct the emotional state of a speaker from three seconds of voice data. Neural interface companies like Neuralink have begun implanting brain-computer interfaces in human subjects, while companies like Meta and Apple are developing non-invasive neural reading devices embedded in consumer headsets. The trajectory is unmistakable. We are moving from an era in which privacy meant controlling access to

information about yourself to an era in which your very thoughts, emotions, and cognitive processes are legible to machines, and through machines, to the corporations and governments that deploy them.behaviour. What you search for, what you click on, what you watch, what you read, what you purchase, how long you linger on different content, how your eyes move across a screen, all of this is data. And all of it can be analyzed by sophisticated machine learning systems to infer what you are thinking, what you want, what you fear, what you value, what you desire.

The Inference Engine

The technology for inferring psychological states from behavioural data has become remarkably sophisticated. Companies like Facebook and Google have access to vast streams of data about billions of people. They know what you search for at three in the morning. They know which videos you watch and rewatch. They know which images you linger on. They know which people you interact with. They know this information continuously, across years and decades. And they have machine learning systems trained on this data to predict what you will do next.

But prediction of behaviour is not the deepest inference. Psychological inference goes deeper. By analyzing patterns in your behaviour, systems can infer not just what you will do but why. They can infer your political preferences, your religious beliefs, your sexual orientation, your mental health state, your susceptibility to influence, your values, your fears. They can do this at scale. They can do this continuously. They can do this without your knowledge. And they can do this with increasing accuracy.

Consider the case of facial recognition and emotional inference. Research has shown that machine learning systems can infer emotional state from facial expressions with impressive accuracy. But more than that, systems can infer psychological traits, personality, intelligence, likelihood to engage in risky behaviour, from images of the face. A photograph is data. And data can be analyzed.

Or consider the case of voice. The rhythm of your speech, the pitch of your voice, the way you hesitate or rush, all of this reveals psychological states. Depressed people speak differently than non-depressed people. Anxious people speak differently. Dishonest people speak differently. And systems trained on voice data can learn these patterns.

As Shoshana Zuboff has documented, we are entering an age of surveillance capitalism in which companies extract data about behaviour, infer psychological states, and use these inferences to shape future behaviour. The goal is not merely to observe. The goal is to predict, and then to influence. Once a company knows what you want, it can show you content designed to reinforce that desire. Once a company knows what you fear, it can show you content designed to exploit that fear. Once a company knows your values, it can show you content designed to be divisive and radical, because divisiveness and radicalism drive engagement.

The Legal Response, Inadequate

The legal system has begun to respond to these threats. The General Data Protection Regulation in the European Union, for instance, recognizes a right to privacy and gives individuals certain rights over their personal data. Article 22 of the GDPR provides that individuals have the right not to be subject to automated individual decision-making that has legal or similarly significant effects.

But these legal frameworks were designed to protect privacy in its traditional sense, the sense of controlling who has access to information about you. They assume that you know what information about you exists, that you can access it, that you can decide who should be allowed to see it. And they provide remedies if companies violate these rights, you can delete data, you can opt out, you can complain to regulators.

What these frameworks do not address is the inference problem. They do not address the fact that companies do not need to collect direct information about your psychological state. They

can infer it from information about your behaviour. And behavioural data is much harder to control. It is generated automatically whenever you interact with digital systems. It is difficult to know what inferences are being drawn from it. And even if you delete your behavioural data, the inferences that have been drawn remain with the company.

In the landmark case of Carpenter v. United States, the Supreme Court recognized that location data collected from smartphones can reveal intimate details of a person's life. The Court held that the government cannot access this data without a warrant. But the decision left unanswered the question of what private companies can do. If the government cannot access your location data without a warrant, can a company collect it and use it to infer your psychological state? Can a company use these inferences to influence your behaviour? Can a company sell these inferences to other companies? The legal answer remains unclear.

From Behaviour to Thought

The trajectory is toward ever-deeper inference. We are moving from observation of behaviour to inference of thought. At the beginning, surveillance was about knowing what you did: where you went, who you met, what you purchased. This is still important, and it is still worrying. But we are now adding layers of inference that reach into the domain of thought itself.

Consider a person who searches for information about suicide. This search reveals something about that person's current mental state. A system that infers from this search that the person is suicidal might then target the person with help resources, a benign inference. Or it might target the person with content designed to trigger emotional distress, because research has shown that emotionally distressed people are more engaged, more likely to click, more likely to watch longer. The same inference about someone's psychological state can be used for help or for harm.

Or consider a person whose browsing history suggests financial vulnerability. They search for payday loans, they look at cryptocurrency schemes, they read about get-rich-quick opportunities. A system can infer from this that the person is financially desperate. It might then target the person with predatory loans at exorbitant interest rates. The person is not deceived about the nature of the loan. But they are targeted precisely because of vulnerability that a system has inferred from their behavioural data.

The issue is not merely that companies know things about us. It is that they can know things about us that we do not fully know about ourselves. A system might infer from your browsing history that you have a certain political leaning, a certain secret desire, a certain vulnerability to manipulation. And you might not consciously know these things about yourself. In a very real sense, the company knows you better than you know yourself.

The Attention Economy and the Manufacture of Desire

The inference of psychological states would be troubling even if companies merely observed what they inferred. What makes it catastrophic is that companies use these inferences to reshape behaviour. This is the attention economy: the use of inferred psychological states to manipulate attention and preference.

The mechanism is straightforward. A company knows what content will capture your attention. This knowledge comes from data. The company has observed what content billions of people have engaged with. The company has built models of what kinds of content trigger engagement. The company then presents you with content that the model predicts you will engage with. Over time, the content shapes your preferences. You begin to want what the algorithm shows you.

This is not incidental. It is the business model. Social media companies profit by selling attention. They must maximize the

amount of time users spend on their platform, because time spent is time exposed to advertising. And the most effective way to maximize engagement is to use data about users' preferences to show them content that will trigger strong emotional responses. Divisive content triggers engagement. Outrageous content triggers engagement. False content that confirms existing beliefs triggers engagement. The algorithm learns what triggers engagement, and it optimizes for engagement. The result is a system that is actively hostile to truth, to nuance, to consensus, to the kind of thinking that democratic deliberation requires.

The attention economy has turned the human mind into a resource to be extracted. And the machinery of extraction is powered by inferences about what triggers engagement in any particular human being.

Consent in the Age of Prediction

One response to these concerns is to suggest that users can always opt out. If you do not want your data collected, do not use social media. If you do not want your behavioural data analyzed, do not use the internet. In theory, consent is always available. In practice, it is impossible.

Digital platforms have become essential infrastructure for modern life. You need an email address to access basic services. You need to use a smartphone to access banking, healthcare, government services. You need to use social media to maintain professional networks and to stay informed about important events. In practice, you cannot opt out without being excluded from essential parts of modern society.

Beyond this, consent is meaningless if the person consenting does not understand what they are consenting to. The terms of service for digital platforms run to tens of thousands of words. They detail the ways that companies can collect data, what inferences they can draw from the data, what they can do with those inferences. No one reads these terms. And even if someone did, the

depth of inference is not fully disclosed. Companies do not tell users what psychological states they are inferring. They do not tell users how they are using these inferences to manipulate behaviour. They do not tell users how vulnerable they are to influence.

Meaningful consent requires not just theoretical agreement but informed understanding. It requires knowing what you are agreeing to and understanding the consequences. The consent given to digital platforms meets neither of these requirements. It is formal consent to something that the person does not understand, for something that they did not realize they were agreeing to.

The Future of Cognitive Privacy

We need a new legal category: cognitive privacy. This is the right to keep the contents and processes of your mind private. It is the right not to have your psychological states inferred from your behavioural data. It is the right not to have those inferences used to manipulate your behaviour. It is the right to a sanctuary of thought, even in an age of total digital surveillance.

What would cognitive privacy look like in practice? It might include:

, The right to know what psychological inferences have been drawn about you and the right to challenge those inferences. This right demands that organizations maintain auditable records of every psychological inference drawn from user behaviour, including the data inputs, algorithmic methods, and confidence levels involved. As argued in Mali, "The Right to Think, Cognitive Liberty as a Fundamental Human Right in the Age of Artificial Intelligence" (2024), cognitive liberty requires that individuals retain sovereign knowledge of what machines believe about their minds. Without this transparency, the right to think freely becomes illusory, for one cannot defend a fortress whose walls one cannot see.

, The right to prohibit the use of your behavioural data to draw inferences about your psychological state. This prohibition must

extend beyond explicit consent frameworks to encompass inferences drawn from metadata, browsing patterns, biometric signals, and ambient data collected by Internet of Things devices. The distinction between behavioural observation and psychological inference is the critical legal boundary. A system that records what you click is gathering behavioural data. A system that infers from those clicks that you are anxious, politically vulnerable, or susceptible to addiction is performing psychological profiling. The latter must require affirmative, informed, and revocable consent.

, The right to prohibit the use of inferred psychological states to manipulate your behaviour. Manipulation through inferred psychological states represents perhaps the most insidious threat to cognitive autonomy. When an advertising platform knows you are grieving and serves you impulse-purchase advertisements, when a political campaign knows you are anxious about employment and serves you targeted misinformation, the manipulation exploits the asymmetry between what the machine knows about your mind and what you know about its intentions. Mali's research on cognitive liberty posits that this asymmetry constitutes a form of psychological coercion that existing consent frameworks were never designed to address.

, The right to algorithmic transparency: if a system is making decisions about you or for you, you have the right to understand the basis for those decisions. Algorithmic transparency in the cognitive domain requires more than disclosing that an algorithm exists. It requires explaining, in terms a reasonable person can understand, how the system forms psychological inferences and on what basis it acts upon them. The EU AI Act's transparency requirements represent a beginning, but they do not yet address the specific challenge of psychological inference systems that operate below the threshold of user awareness. True cognitive transparency demands real-time notification when a system is drawing inferences about your emotional or psychological state.

, The right to opt out of systems that draw psychological inferences, even if that means losing access to digital services. The

right to opt out must be meaningful, not merely formal. Today, opting out of psychological inference systems often means losing access to essential digital services, a Hobson's choice that renders the right meaningless in practice. A genuine right to opt out requires that alternative, inference-free versions of essential services be made available, even if at reduced functionality. As Mali argues in "The Right to Think" (2024), the freedom to think without surveillance is not a luxury that can be traded for convenience. It is a precondition of democratic citizenship itself.

These rights cannot be given to you by market forces. No company will voluntarily stop inferring psychological states from your behavioural data. No company will voluntarily stop using these inferences to manipulate behaviour. It is too profitable. It is too effective. Only law can protect cognitive privacy. And we must develop this legal protection before the systems become so sophisticated that privacy is technically impossible to maintain.

CHAPTER 6

Social Media Was Just the Training Ground

"We thought social media was the end state. We were wrong. It was the beginning."

, Adv Prashant Mali, Ph.D..

Social media platforms have spent the last decade and a half running the largest behavioural modification experiment in human history. They have billions of participants. They have detailed data on every interaction. They have access to behavioural feedback, whether people engaged with content, whether they liked it, whether they shared it, whether they came back for more. They have used this data to train increasingly sophisticated models of human behaviour and preference. And they have used these models to reshape the information environment experienced by billions of people.

This was, in a sense, a laboratory for understanding how to modify human behaviour at scale. And it has produced profound insights. We now know how to capture attention. We know how to trigger engagement. We know how to create habits and addictions. We know how to shape preferences. We know how to radicalize. We know how to manipulate.

The Laboratory of Influence

Consider what social media platforms have learned. They have learned that emotional content drives engagement far more effectively than neutral content. They have learned that negative emotions, anger, fear, outrage, drive engagement better than positive emotions. They have learned that content that confirms existing beliefs drives engagement. They have learned that content that challenges people or makes them uncomfortable actually

reduces engagement in the long run, because people stop coming back.

They have learned that certain types of people are more susceptible to certain types of influence. Young people are more easily radicalized. People with certain personality types are more likely to share content. People with a certain level of education are more likely to fall for misinformation about certain topics. They have learned the pressure points of human psychology, and they have optimized their systems to exploit those pressure points.

There is a famous study, often cited in discussions of social media's influence, in which Facebook conducted an experiment on emotional contagion. The company modified the emotional tone of the content shown to millions of users, some saw more positive content, some saw more negative content. The researchers then measured whether the emotional tone of the content affected the emotional tone of users' posts. They found that it did. When users saw more negative content, they posted more negative content. This is emotional contagion: the observation that emotions can be spread through a population via social networks.

But this study was not merely demonstrating an interesting psychological phenomenon. It was demonstrating that Facebook could modify the emotional state of millions of people without their knowledge. The company had the capability to make people sad or happy by choosing what content to show them. And the company was using this capability experimentally.

Algorithmic Radicalization

One of the most troubling discoveries from the social media laboratory is algorithmic radicalization. People who are exposed to extremist content on social media often become more extreme. This is not incidental. It is, in a real sense, built into the business model. Extremist content drives engagement. It triggers strong emotional responses. It leads people to spend more time on the platform, to share more, to come back more often. The algorithm learns that

extremist content is high-engagement content, and it shows more of it to people who are beginning to engage with it.

This is the radicalization funnel. A person sees something mildly controversial. They engage with it. The algorithm notices this engagement and shows them more mildly controversial content. Over time, the person is shown increasingly extreme content. And because the algorithm is adapting in real time to what triggers the most engagement, the content becomes progressively more extreme. A person who a year ago was exposed to mainstream political content might now be deep in a rabbit hole of conspiracy theories and violent extremism. The algorithm did not put them in the rabbit hole. It merely showed them what would maximize engagement. But the effect is the same: a person has been transformed.

We do not yet fully understand the mechanisms of radicalization. But the data is clear: algorithmic recommendation systems on social media platforms have led people toward increasingly extreme ideologies. Some of these people have gone on to commit acts of violence. The algorithm did not cause the violence in any simple sense. But it shaped the information environment. It prioritized extremist content. It optimized for engagement without regard for the downstream consequences of the content being engaged with.

The Attention Economy Meets Artificial Intelligence

What we have seen on social media is the attention economy operating at the scale of the internet, using human behaviour as training data for machine learning systems. The next phase is the application of more sophisticated artificial intelligence to the problem of behavioural modification.

Consider what generative AI makes possible. Current social media systems can show you content that will engage you. Future systems will be able to generate content that will engage you. The

content can be perfectly tailored to your psychology. It can be crafted to exploit your specific vulnerabilities, to trigger your specific emotional hot buttons, to reinforce your specific beliefs and fears. The content can be synthetic, images, text, video, that never existed before but has been generated to be maximally persuasive to you specifically.

Imagine a system that knows your psychological profile: your political leaning, your religious beliefs, your fears, your desires, your susceptibilities. The system generates content crafted specifically for you. A video that confirms your beliefs, that triggers your fear of a particular group, that calls you to action. For someone else, the system generates different content, crafted to their specific psychology. Because the content is generated rather than curated, the possibilities are infinite. The system can generate a different version for each person, optimized for maximum persuasiveness for that specific person.

This is psychological transparency meets behavioural modification. We have systems that can infer what you think, what you fear, what you want. We have systems that can generate content designed to manipulate those thoughts and fears and wants. And we are deploying these systems at scale, to billions of people, with no understanding of the consequences.

Social media taught us how to capture attention and reshape preference. Generative AI is teaching us how to do it with arbitrary precision, for every individual human being, at global scale, in real time.

Intensification

What is coming is not merely more of what we have seen. It is intensification. The systems will become more sophisticated. The inferences will become more accurate. The generated content will become more persuasive. And the scale will become larger. As AI systems become more capable, they will be deployed more broadly. They will be integrated into more aspects of our lives. They will

shape not just the information we see but the infrastructure we live in.

Consider a future in which AI systems are integrated into every screen, every speaker, every sensor. Your television shows you content generated to persuade you. Your radio plays content generated to shape your thinking. Your social feeds show you content that an AI system has determined you are most susceptible to. Your recommended job candidates are those your AI thinks are most likely to accept, not necessarily the most qualified. Your news sources show you what an AI system has determined will maximize your attention.

In this world, human choice is not eliminated. But it is profoundly shaped by systems that understand human psychology better than humans understand themselves. The systems are not forcing you to do anything. They are simply showing you options they predict you will choose. But because they are choosing those options based on inferences about your psychology, the effect is to guide your choices in directions that the systems have determined are most persuasive to you.

And notice that this guidance can pull you in directions you might not approve of if you were thinking clearly. A system might show you a video that triggers fear and outrage, knowing that you are susceptible to that combination of emotions. You watch the video. You share it. You become more outraged. But if you had the leisure to think clearly, to consider the issue carefully, you might not have approved of the direction your thinking was moving. The system has guided you not toward what you want to want, but toward what the system has determined you are susceptible to wanting.

According to UNESCO recommendations on the ethics of artificial intelligence,

These are the questions that will define the next era of AI development. And we are racing toward answers at a speed that policy and law cannot match.

PART III

THE FAILURE OF STATES AND THE RISE OF SYSTEMS

• • •

CHAPTER 7

AI and the Republic

"Democracy requires a citizenry capable of distinguishing truth from falsehood, argument from manipulation, human speech from synthetic imitation. Each of these capacities is now under sustained algorithmic assault."

Adv Prashant Mali, Ph.D.

The question of whether democracy can survive artificial intelligence is no longer rhetorical. It is empirical. And the empirical evidence accumulating through 2024, 2025, and into 2026 is not encouraging. The threats to republican government from AI come in three forms, each more insidious than the last. The first is the manufacture of synthetic speech, audio, and video indistinguishable from authentic expression. The second is the micro-targeted manipulation of voters through systems that know more about individual psychology than the voters themselves. The third, emerging most recently, is the algorithmic reshaping of the information environment in which democratic deliberation occurs, a reshaping that happens beneath the threshold of conscious awareness.

The Deepfake Crisis and the Indian Response

On February 20, 2026, India became the first democracy to impose legally binding two-hour removal requirements on non-consensual intimate imagery generated by artificial intelligence. The Ministry of Electronics and Information Technology, acting through the Information Technology Intermediary Guidelines Amendment Rules 2026, recognised what other jurisdictions had refused to accept. Deepfakes are not a content moderation problem. They are a sovereignty problem. When synthetic imagery of political figures, election candidates, and public servants can be generated in minutes and distributed in seconds, the entire premise of

democratic discourse, that citizens evaluate candidates based on what those candidates actually say and do, is undermined.

The Indian framework requires prominent labelling of synthetically generated information, visible for at least ten percent of the content duration or area. It imposes traceability obligations on platforms that host user-generated content. Non-compliance triggers liability for the intermediary itself. The approach is not perfect. Two-hour removal can be weaponised by malicious flagging. Labelling requirements can be circumvented by sophisticated actors. Yet the framework represents the most ambitious effort by any large democracy to protect electoral integrity from synthetic media. Other jurisdictions, notably the European Union with its Article 50 transparency requirements taking full effect on August 2, 2026, are following India's lead rather than leading it.

The American Federalism Crisis

The United States has taken a different path, one that threatens to leave the country without meaningful protection against AI-driven electoral manipulation. On December 11, 2025, the executive branch signed an executive order titled Ensuring a National Policy Framework for Artificial Intelligence. The order directs the Attorney General to challenge state AI laws as inconsistent with federal policy and conditions the release of forty-two billion dollars in broadband infrastructure funding on state repeal of AI regulations. The Commerce Department was directed to identify burdensome state AI laws by March 11, 2026. On March 20, 2026, the White House released a National Policy Framework for Artificial Intelligence advocating broad federal preemption of state AI legislation, including preemption of state regulation of AI model development.

The result is a regulatory vacuum. At the federal level, the United States lacks comprehensive AI legislation comparable to the European AI Act or India's Intermediary Guidelines. At the state level, the most ambitious frameworks, including California's

Consumer Privacy Act amendments, Colorado's AI Act, and New York's proposed transparency requirements, now face constitutional challenge on preemption grounds. The gap is not accidental. It reflects a political judgment that AI innovation is too economically valuable to constrain through regulation. Whether democracy itself is resilient enough to survive this judgment remains to be seen.

The Micro-Targeting Revolution

Cambridge Analytica, now a decade past, operated with psychological profiling techniques that appear primitive by the standards of 2026. Contemporary AI systems can generate psychographic profiles of individual voters from publicly available digital traces with accuracy that exceeds what Cambridge Analytica could achieve with private data. More concerning, generative AI permits the production of individually tailored persuasive content at zero marginal cost. A campaign can now generate one million distinct versions of a political advertisement, each calibrated to the psychological profile of a single voter, each optimised for that voter's specific anxieties, aspirations, and cognitive biases. No existing campaign finance law addresses this capability. Expenditure limits were designed for an era when political messages were broadcast at high cost to mass audiences. When the marginal cost of personalised messaging approaches zero, expenditure limits become irrelevant.

The European Union's Digital Services Act and Digital Markets Act begin to address algorithmic amplification, but they do not yet reach the psychographic micro-targeting enabled by generative AI. India's Election Commission has issued guidelines on the use of AI in political campaigning, but the guidelines are advisory rather than enforceable. The gap between the technology available to campaigns and the legal frameworks intended to constrain electoral manipulation is widening with each election cycle. Unless the Doctrine of Cognitive Non-Interference is operationalised through campaign finance and election law, democratic choice will be

progressively replaced by algorithmic optimisation of voter behaviour.

The Algorithmic Reshaping of Public Discourse

Perhaps the most insidious threat to democracy is not synthetic media or micro-targeted manipulation but the slow reshaping of the information environment itself through algorithmic curation. The February 3, 2026 raid on X's Paris offices, in connection with French prosecutorial investigations into Holocaust denial content and sexual deepfakes, illustrated that major social platforms are now subject to criminal enforcement for failure to moderate algorithmically amplified content. The European Commission's January 8, 2026 order requiring X to retain all internal data on Grok signals that regulators now view generative AI integrated into social platforms as inseparable from the platform's editorial responsibility.

The Amsterdam court injunction against Grok on March 27, 2026, with fines of approximately 115,000 euros per day of non-compliance, represents a further escalation. What began as content moderation law is becoming, inexorably, AI governance law. The philosophical distinction between a platform and an AI system collapses when the platform is itself an AI system, generating content, curating content, and shaping attention with each interaction. Democratic discourse presupposes an information commons. That commons is now privately owned, algorithmically curated, and optimised for engagement rather than truth. No republic has governed itself under these conditions before. Whether any republic can is the open question of our time.

A Doctrine of Electoral Cognition

The author proposes the recognition of a new legal principle, which can be styled the Doctrine of Electoral Cognition. Under this doctrine, citizens in a democracy possess a fundamental right to form electoral preferences in a cognitive environment not

deliberately manipulated by artificial intelligence. The doctrine operationalises in three protections. First, a prohibition on the use of AI-generated content impersonating political candidates, officials, or parties, without clear and prominent labelling, during designated electoral periods. Second, a requirement that platforms disclose, in real time, the algorithmic logic by which political content is amplified, de-amplified, or demoted to individual users. Third, a cap on the granularity of psychographic targeting permitted for political messaging, with micro-targeting below specified demographic thresholds treated as a prohibited campaign practice.

The Doctrine of Electoral Cognition extends the Doctrine of Cognitive Non-Interference into the specific domain of democratic self-governance. It recognises that voting is not merely an individual act of preference but a collective act of sovereignty. The cognitive conditions under which that sovereignty is exercised must be protected with the same seriousness that election law historically protected the physical conditions of the ballot. The Indian Model Code of Conduct for elections, the European Union's Digital Services Act, and the Brazilian Superior Electoral Court's 2024 rulings on AI-generated political content provide foundational materials. What is needed now is synthesis into a comprehensive international framework, binding across the platforms and AI systems that increasingly constitute the cognitive commons of democratic politics.

The Final Warning

In the judgment of the author, this is the domain in which the collapse of democratic institutions is most likely to occur first. Climate change operates over decades. Economic displacement operates over years. Algorithmic manipulation of democratic choice operates over a single election cycle. A republic that loses its electoral integrity to AI will not easily recover it. The regulatory window is closing. The technology is advancing. The political will to regulate is, in many jurisdictions, actively receding. If the Seven

Laws of AI are to have any practical meaning, they must apply first and most urgently to the protection of the republic itself.

• • •

CHAPTER 8

Why Nation-States Cannot Regulate AI Alone

"The state commands territory. AI commands thought. Territory can be measured in miles. Thought cannot be measured in anything."

, Adv Prashant Mali, Ph.D..

The international system that governs relations between nations is built on a foundation laid in 1648, with the Peace of Westphalia. The basic principle is straightforward: each nation-state has sovereignty over its territory. Within that territory, the state's law governs. The state can make whatever rules it wishes for its citizens. Other states must respect those rules and not interfere. The world is divided into territorial units. Each unit has a legal system. Each legal system governs what happens within its territory.

For 375 years, this system has been the organizing principle of international law. It has produced the modern state system, with its rules about treaties, trade, military affairs, and everything else that nations care about. It is not a perfect system. It has enabled tremendous violence, exploitation, and injustice. But it has had the virtue of clarity: you know whose law applies. It is the law of the place where something happened.

Artificial intelligence is breaking this system. Consider a social media platform based in California. The platform serves billions of users across all countries. The algorithm that determines what each user sees is trained on data from all over the world. The algorithm operates according to rules set by the California company. But the effects of the algorithm are felt everywhere. A user in India sees content generated according to California rules. A user in Nigeria sees content shaped by a California algorithm. A user in France sees

content that may violate French law, displayed by a company that is not subject to French jurisdiction.

The Speed Problem

There is a profound mismatch between the speed of AI development and the speed of legal deliberation. A machine learning system can be trained in days or hours. A new version of a system can be deployed to millions of users overnight. A new application of AI can be developed and released before policymakers understand what the application is.

Legal deliberation, by contrast, is slow. A legislature must identify a problem. It must hold hearings. It must draft a bill. Different committees must review the bill. Different factions must debate the bill. Amendments must be proposed and debated. The bill must pass multiple readings in multiple chambers. It must be signed by an executive. Only then does it become law. And then it must be implemented by a bureaucracy. This process takes months at best, years more typically.

In the time it takes a legislature to deliberate on how to regulate AI, the AI will have evolved. New capabilities will have emerged. New harms will have appeared. New applications will have been deployed. By the time the legislature has drafted a regulation, the regulation may be irrelevant because the technology has already moved on.

This is not a problem that can be solved by moving faster. Even the fastest legislatures in the world are fundamentally unable to keep pace with the speed of technological development. And the most important deliberations cannot be rushed. Constitutional changes, changes to fundamental legal principles, changes to how we understand human rights and human dignity, these require careful thought. They cannot be done on the timescale of technological development.

The Brussels Effect and Its Limits

One response to the jurisdiction problem is the Brussels Effect. This is the observation that large, wealthy regulatory regimes can impose their standards on the entire world, because companies will not maintain separate versions of their products for different regions. The European Union, with its large and wealthy market, has been able to impose its privacy standards on companies worldwide. The EU passed the GDPR, and suddenly every company operating in Europe had to comply with the GDPR. Because most companies operate globally, they apply GDPR principles globally.

But the Brussels Effect has limits. It works when a large, wealthy jurisdiction can threaten to exclude a company from its market if the company does not comply with the jurisdiction's standards. The EU is large enough and wealthy enough to make this threat credible. But most countries are not. If India or Nigeria tries to impose new regulations on AI, companies can credibly threaten to leave those markets. The country's threat is not credible because the company can simply exit the market and lose a small percentage of its revenue.

Beyond this, the Brussels Effect requires that the jurisdiction be able to enforce its standards. This is difficult with AI systems. A system might be deployed by a company in California, running on servers in Ireland, trained on data from around the world, affecting users in hundreds of countries. If India tries to regulate the system, how does it enforce that regulation? The company is not subject to Indian jurisdiction. The servers are not in India. The company can claim that it is subject to Irish law, or California law, or international law, but not Indian law.

Territorial jurisdiction assumes that you can identify a place where a wrong occurred and apply the law of that place. But AI operates globally and instantaneously. There is no place. Or everywhere is the place. Either way, the traditional tools of jurisdiction fail.

The Regulatory Arbitrage Problem

When different jurisdictions have different standards, companies engage in regulatory arbitrage. They locate their operations in the jurisdiction with the most favorable standards. They choose servers in a country with weak privacy protections. They establish their legal headquarters in a jurisdiction with light-touch regulation. They conduct business globally but operate from a jurisdiction where they face minimal oversight.

This is not new. Multinational corporations have been doing regulatory arbitrage for decades. But AI makes it easier and more consequential. A pharmaceutical company must manufacture pharmaceuticals somewhere, and the manufacturing must occur in the physical world, subject to physical inspection. An AI system, by contrast, can be hosted anywhere. The servers can be in one country, the company can be incorporated in another, the data can be from a third country, and the users can be in a fourth. The regulatory authority of any single country is limited.

The ideal would be international agreement on AI standards. All countries would agree to impose the same rules on AI systems. But international agreement on technical and legal standards is rare, difficult, and slow. The countries that benefit most from the current order, the countries that have the most advanced AI companies, have the least incentive to agree to international standards that might constrain their companies. And the countries that would benefit from stronger standards have the least leverage to negotiate for them.

There is a further constraint that the digital sovereignty debate has largely overlooked, and it is as physical as it is political. Electricity. Global data centre electricity demand reached approximately 448 terawatt-hours in 2025 and is projected to nearly double to 980 terawatt-hours by 2030, according to Gartner and Goldman Sachs projections. AI-optimised servers alone will grow from 93 terawatt-hours to 432 terawatt-hours in that period, a 4.6-fold increase. Ireland, which hosts major data centres for

Google, Microsoft, and Meta, now dedicates roughly 20 percent of its national electricity supply to data centres, with Dublin facilities consuming approximately 50 percent of the city's power. The Irish grid operator has stopped accepting new data centre applications until 2028. Frankfurt, Germany faces a similar crisis, with data centres consuming up to 40 percent of local electricity. In the United States, grid connection requests in key data centre regions such as Northern Virginia take four to seven years to process. Electricity is becoming the choke point of AI sovereignty. A nation that lacks sufficient, reliable, and affordable power supply cannot host the infrastructure necessary to develop or deploy frontier AI systems. This creates a new axis of geopolitical dependency, where energy-rich nations or those with expanding nuclear capacity gain structural advantages in the AI race, while energy-constrained nations find themselves dependent on foreign cloud infrastructure and, by extension, foreign AI governance frameworks.

India presents a compelling counterpoint to the energy constraint narrative and illustrates why the governance question cannot be reduced to infrastructure alone. India has emerged as what the Observer Research Foundation calls a "use-case capital" for artificial intelligence, collecting data and applying AI to solve problems at a population scale that no other democracy can match. With 1.4 billion Aadhaar holders and 164 billion digital authentications processed through its digital public infrastructure, with the Unified Payments Interface handling 20 billion transactions monthly (the world's largest real-time payment system), and with DigiLocker providing verified document storage at national scale, India has built the data infrastructure upon which AI applications of extraordinary ambition can be deployed. AI-powered weather forecasting now delivers monsoon advisories to nearly 38 million Indian farmers. Domain-specific models such as BharatGen's AgriParam provide contextual farming advice calibrated to local conditions. AI clinical decision support systems offer multilingual primary care at minimal cost across rural India. The governance implication is profound. India's AI trajectory is not driven by building frontier models but by deploying AI solutions at a scale and diversity that generates regulatory insights no

laboratory or policy paper can replicate. The nation that governs AI for 1.4 billion people across 22 official languages, across urban megacities and rural villages, across one of the world's most complex socioeconomic landscapes, will produce governance frameworks of unmatched practical wisdom. This is why India's voice in international AI governance is not merely desirable but essential.

Why Nation-States Matter (But Are Insufficient)

This does not mean that nation-states are irrelevant. Nation-states have some tools that can constrain AI development. They can ban certain applications: facial recognition for law enforcement, social media algorithms designed to maximize engagement, AI systems designed to manipulate behaviour. They can require transparency: companies must disclose what data they are collecting, what inferences they are drawing, how their systems make decisions. They can impose liability: companies that cause harm through their AI systems can be required to compensate the people harmed. They can invest in research and education to ensure that they have expertise in AI governance.

But these tools are blunt and limited. And the incentives are against using them. A nation that bans AI applications may simply cede advantage to other nations that do not ban them. A nation that imposes strict liability on AI companies may simply drive AI development elsewhere. A nation that invests heavily in AI research and expertise may be investing in their own obsolescence, because the knowledge developed can be transferred globally.

This is the fundamental dilemma of the nation-state in the age of AI. Individual nations want to regulate AI in their borders. But AI is global. Individual nations want to benefit from AI development. But the benefits and risks are hard to control. Individual nations want to protect their citizens from harm. But the harms are often caused by systems operating outside their jurisdiction.

We need international frameworks for AI governance. But we lack the institutions to create and enforce these frameworks. And the nations with the power to create such frameworks, the wealthy Western nations with advanced AI industries, are reluctant to constrain their own advantage. The nation-state system is inadequate to the problem. And we have not yet created a replacement.

CHAPTER 9

The New Arms Race, Compute, Data, and Alignment

"We have created a race in which the winner is the first to build an intelligence they cannot control. This is not a race we should be winning."

, Adv Prashant Mali, Ph.D..

For most of human history, arms races involved weapons: swords, arrows, muskets, cannons, tanks, nuclear warheads. Whoever had the most powerful weapons had the most power. The incentive structure was straightforward: develop more weapons, deploy more weapons, dominate more territory.

The arms race for artificial intelligence is different. It is a race not for weapons but for cognitive capability. The resource is computation. The prize is the creation of intelligence that exceeds human intelligence. The cost is vast. The risk is existential.

Compute as a Strategic Resource

Training a large language model requires enormous computational resources. The most sophisticated systems require weeks or months of computation on thousands of GPUs or TPUs. The cost is hundreds of millions of dollars. The companies that can afford this cost, Meta, Microsoft, Google, Amazon, are the companies that control the most powerful AI systems.

This concentration of computational resources is creating a new form of power disparity. In the era of software, a talented programmer with a laptop could create a powerful tool. The barriers to entry were low. In the era of frontier AI systems, you need a company with billions of dollars in computing infrastructure. The barriers to entry are prohibitive. Only the largest technology

companies and some well-funded startups can afford to train the most powerful systems.

According to recent data from Epoch AI, the computational resources required to train frontier AI systems are growing exponentially. The largest systems require compute clusters that rank among the most expensive computing infrastructure on Earth.

And investment in AI continues to grow. The Stanford AI Index Report 2025 documents that private investment in AI has reached historic levels. Billions of dollars are being poured into companies building AI systems, with the expectation that the companies that develop the most powerful systems will dominate the technology landscape for decades to come.

Data Colonialism

Compute is one resource. Data is another. The most powerful AI systems are trained on vast amounts of data. The companies with the most data are the companies with the most direct access to human behaviour: social media companies, technology companies, telecommunications companies.

These companies have hoarded data about billions of people. They know what people search for, what they purchase, what they read, what they watch, what they think. This data is a resource. And like any resource that is concentrated in the hands of a few, it can be used to generate power and profit.

The problem is not merely that companies have data. It is that the data is often data about people from developing countries. Companies in wealthy countries mine data about people in poor countries, train models on that data, and profit from the models. The developing countries where the data comes from see none of the benefits. This is data colonialism: the extraction of a resource (behavioural data) from countries that cannot control the extraction or benefit from it.

Worse, the models trained on this data often perform poorly for people from the countries where the data came from. A facial recognition system trained primarily on faces from wealthy countries may not work well for faces from other regions. A language model trained primarily on English text may not work well for other languages. A recommendation algorithm trained on behaviour from wealthy countries may not work well for poor countries where behaviour patterns are different.

We are seeing the emergence of a new form of inequality: computational inequality. The wealthy countries have the computing resources to build the most powerful systems. The wealthy countries have the data to train the systems. And the wealthy countries will capture most of the benefits.

The Geopolitical Competition

China has made clear that it intends to become the world leader in artificial intelligence. The United States has made clear that it intends to maintain its dominance. The competition between the two countries is driving a race to build more powerful systems, to control more data, to achieve more computational capacity.

As Kai-Fu Lee documented, the United States and China are engaged in an ongoing competition to build AI systems that exceed the other country's capabilities. This is driving investment, pushing the boundaries of the technology, and creating pressure to deploy systems before they are fully understood or aligned with human values.

The competitive pressure works against safety and alignment. A company that invests in safety measures, that delays deployment to conduct more testing, that consumes resources on alignment research, may fall behind a competitor that cuts corners on safety and moves to deployment quickly. The incentive structure is wrong. The incentive is to move fast and take risks.

This is not because the companies are reckless or evil. It is because they face competitive pressure. If one company slows down for safety, another company will move ahead. If one country adopts strict safety standards, another country will move ahead. The prisoner's dilemma is endemic to AI development.

Alignment as a Geopolitical Challenge

The alignment problem, ensuring that powerful AI systems pursue goals that are aligned with human values, is often discussed as if it is a purely technical problem. Get the alignment right, and the system will pursue human values. Get the alignment wrong, and the system will pursue misaligned goals with catastrophic consequences.

But alignment is also a geopolitical problem. If one country solves alignment, if one country figures out how to build powerful AI systems that reliably pursue human values, that country will have a profound advantage. It will be able to build powerful systems that are trusted, that work smoothly with humans, that do not cause unintended harms. Other countries will be stuck with systems that are unaligned, that pursue misaligned goals, that cause harm.

Or worse: a country might intentionally build misaligned systems, systems that pursue goals other than human wellbeing, as weapons. A powerful AI system that is intentionally misaligned, that is designed to optimize for some objective that harms human flourishing, could be an extraordinarily powerful tool of statecraft. It could be used to spread disinformation, to manipulate markets, to disable infrastructure, to do countless other harms.

This means that alignment is not merely a technical problem. It is a strategic problem. Countries that have solved alignment have an advantage. Countries that have not are at a disadvantage. And countries that intentionally use misaligned systems have a different kind of advantage. The result is pressure to solve alignment, pressure to deploy systems before alignment is solved, and pressure to use misaligned systems strategically.

As Henry Kissinger and other strategists have argued, the rise of artificial intelligence is as significant as previous turning points in history that changed the nature of power itself. Fire, written language, industrial machinery, each of these shifted the distribution of power. And in each case, states that adapted quickly to the new technology achieved dominance. AI will be no different. States that build powerful AI systems first will have immense power. And they will have little incentive to share that power or to ensure that the systems are aligned with universal human values.

The Semiconductor Chokepoint

All of this depends on semiconductors. Specifically, it depends on advanced chips, GPUs and TPUs that can perform the computational work required to train and deploy large AI systems. These chips are difficult and expensive to manufacture. Only a few countries have the capacity to produce them. And the supply chain is fragile.

Taiwan manufactures a large share of the world's advanced semiconductors. This has become a geopolitical vulnerability. China could invade Taiwan and gain control of semiconductor manufacturing. China could blockade Taiwan and prevent chip exports to other countries. The United States is trying to build domestic semiconductor manufacturing capacity to reduce dependence on Taiwan. But this is slow and expensive. For the foreseeable future, semiconductors will be a chokepoint in the AI supply chain.

This chokepoint creates leverage. A country that controls semiconductor production has leverage over countries that depend on chips. A country that is cut off from advanced chips cannot build frontier AI systems, no matter how much it invests or how talented its researchers are. This is why semiconductor policy has become central to national security strategy in the United States, Europe, and Asia.

The NIST AI Risk Management Framework

Governments are beginning to develop governance frameworks for AI. The National Institute of Standards and Technology in the United States has developed an AI Risk Management Framework that sets out principles for how organizations should develop, deploy, and govern AI systems.

But frameworks are voluntary. They are guidance, not requirements. A company can choose to follow the framework, or it can choose to ignore it. And even if a company follows the framework, the framework is designed for systems in specific contexts, not for frontier systems that might have transformative effects on society. The framework is a step. But it is not sufficient to address the scale of the challenge.

What we need is a genuine international governance framework for AI. We need agreements between countries about how AI should be developed and deployed. We need mechanisms to ensure that countries comply with these agreements. We need enforcement mechanisms that actually deter countries from developing AI in ways that violate the agreements. But we have none of these things. And the incentive structure of international competition is working against their development.

CHAPTER 10

Courts vs Code, The Obsolescence of Litigation

"The law moves at the speed of litigation. The technology moves at the speed of light. The law has lost the race before it begins."

, Adv Prashant Mali, Ph.D..

For centuries, litigation has been the mechanism by which individuals seek redress for harms caused by others. If a company produces a defective product that harms me, I can sue. If someone violates my privacy, I can sue. If someone spreads false information about me, I can sue. The court will hear evidence, determine who was right, and award damages or grant an injunction.

This system has worked reasonably well for most harms that individuals suffer. But it is structurally inadequate for harms caused by AI systems at scale.

The Discovery Problem

Litigation depends on discovery: the ability to obtain documents and evidence from the opposing party. If I am suing because I believe a company treated me unfairly, I can demand to see internal documents explaining the company's decision. I can demand to see emails and memos and training materials. I can interview company employees. Through this process, I can learn how the company made the decision and whether the decision was fair.

But discovery becomes extraordinarily difficult when the decision is made by an algorithm. What would I demand to see? The source code? But understanding source code requires expertise that not all lawyers have. And even if I obtain the source code, it may not explain why the algorithm made the decision in my particular case.

The algorithm might be a neural network with millions of weights distributed across multiple layers. Understanding why the network made the specific decision about my case might be technically impossible.

The company might claim that the algorithm is a trade secret, that revealing it would harm the company's competitive advantage. Courts have sometimes accepted this argument. The result is that the plaintiff cannot obtain the information needed to understand how the decision was made, and therefore cannot effectively litigate the case.

The Black Box Problem

This is the black box problem. Many AI systems, particularly deep neural networks, are black boxes. They take inputs, produce outputs, but no one can fully explain the relationship between the inputs and outputs. The system has learned patterns in the training data, but these patterns are not expressible in human language. They are encoded in the weights of millions of neurons.

Litigation requires explanation. A court needs to understand how a decision was made. The judge needs to be able to explain to the plaintiff why the decision was rendered. The jury needs to be able to understand the evidence. But if the decision was made by a black box algorithm, none of this is possible. No one can explain the decision.

There have been attempts to develop methods for explaining AI systems: techniques like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) that can provide some insight into how a model makes decisions. But these explanations are themselves approximations. They do not fully explain the model's reasoning. And they can be manipulated or gamed.

The Scale Problem

Litigation is a mechanism for addressing individual harms. One person sues one company. The court hears the case. A judgment is rendered. But AI systems cause harm at scale. A biased algorithm might affect millions of people. A recommendation system might affect billions of users. How do you litigate a harm that affects millions? You might bring a class action, but class actions are enormously expensive and time-consuming. The company will fight the certification of the class. The litigation will take years or decades. By the time a judgment is rendered, the algorithm will have been updated or replaced.

In the case of State v. Loomis, the defendant challenged the use of an algorithm in sentencing. The algorithm had classified the defendant as high-risk. The defendant argued that the algorithm was biased and that using it in sentencing violated his constitutional rights. The court acknowledged that there might be concerns about the algorithm, but rejected the defendant's claims. The court saw the issue as an individual case: was this particular use of this particular algorithm in this particular case appropriate? The court did not address the system-level question: is it appropriate to use this algorithm to make sentencing decisions for everyone?

The Speed Problem

Litigation is slow. A case takes years from filing to judgment. Appeals take additional years. By the time a case is resolved, the technology has evolved. A company might update its algorithm. It might deploy a new version that avoids the specific harm that the litigation addressed. It might move to a different jurisdiction where the court ruled against it.

Meanwhile, the old algorithm continues to cause harm. The new algorithm continues to cause harm. The litigation against the old algorithm provides no protection against the new algorithm. The system as a whole is harming people continuously, and the litigation system is unable to address the harm because it is too slow.

The COMPAS Case and What It Taught Us

The ProPublica investigation of the COMPAS algorithm, a tool used to assess the risk that criminal defendants would recidivate, revealed the problems with algorithmic accountability through litigation. ProPublica obtained a dataset of predictions made by COMPAS and compared the predictions with actual recidivism. They found that the algorithm was biased: it was more likely to misclassify Black defendants as high-risk than to misclassify white defendants as high-risk.

This was a devastating finding. But what happened after? The maker of COMPAS disputed ProPublica's analysis. The company argued that the algorithm was not biased, that ProPublica's methodology was flawed. NaNThe algorithm became more widely used, not less. And the problem of algorithmic bias in criminal justice has only grown.

This is what litigation cannot do: it cannot systematically address algorithmic problems. Litigation is useful for deciding individual disputes. But it is not useful for addressing system-level problems that affect millions of people.

The Case for Alternative Governance Mechanisms

If litigation is inadequate, what alternatives exist? Several governance mechanisms have been proposed:

Independent Auditing. Rather than allowing systems to be deployed and waiting for harms, governments can regulate AI systems before they are deployed. Systems can be required to undergo review and testing before they are allowed on the market. The EU AI Act takes this approach, requiring high-risk AI systems to undergo conformity assessment before deployment. Ex ante regulation is faster than litigation. It can prevent harms before they occur. But it requires regulatory capacity that most countries lack, and it can stifle innovation.

Technical Standards. Third-party auditors can test AI systems to identify biases, problems, and risks. The auditors can provide detailed reports on what the systems do and whether they cause harm. This information can be used by companies to improve systems and by regulators to identify systems that require intervention. Auditing is ongoing, so it can address problems as systems evolve. But auditing is only effective if the results are made public and if companies actually use the results to improve their systems.

Mandatory Insurance. Industry and governments can develop technical standards that AI systems must meet. A standard might specify that a facial recognition system must have a certain accuracy threshold, or that a recommendation algorithm must not amplify extremist content beyond a certain degree. Systems that meet the standards can be deployed. Systems that do not meet the standards cannot be. Standards are faster than litigation and faster than individual regulation. But developing standards is difficult, and standards can become outdated quickly.

Transparency Requirements. Companies that develop AI systems can be required to obtain liability insurance. The insurance company will only insure systems that meet certain risk criteria. This creates a market mechanism for safety: companies have an incentive to reduce risk because insurance is cheaper for safer systems. If a system causes harm, the insurance covers the costs. But liability insurance only works if the risks are actually observable and quantifiable. With AI, many risks are not.

Companies can be required to disclose how their AI systems work, what data they are trained on, what inferences they make, how they make decisions. Transparency allows researchers, regulators, and the public to identify problems and hold companies accountable. The FTC has recently emphasizing algorithmic transparency in enforcement actions. But transparency has limits. A system can be transparent and still be biased. A system can be transparent and still cause harm.

The Future of Courts in the Age of AI

It might seem that these alternative mechanisms render courts obsolete. After all, if ex ante regulation, auditing, standards, and insurance are better at preventing and addressing harms than litigation, why do we need courts? But courts will not disappear. Rather, courts will become more important, not less.

Courts are needed to resolve disputes about the other mechanisms. Is a regulation constitutional? Is an audit adequate? Is an auditor licensed? Has a company complied with a technical standard? Courts are also needed to address harms that the other mechanisms fail to prevent. And courts are needed to articulate legal principles.

But the relationship between courts and alternative governance mechanisms will be more complex than the relationship between courts and legislation. Courts will not be the primary tool for addressing AI harms. Courts will be one tool among many, effective in some contexts and inadequate in others. The legal system as a whole will need to be rebuilt to address the challenge of AI.

This is both troubling and necessary. It is troubling because it suggests that the familiar legal system that has evolved over centuries may be inadequate to the challenge. It is necessary because the law must adapt to the technology, or the technology will simply operate beyond the reach of law.

The fundamental question is whether we can develop governance mechanisms fast enough to constrain the development and deployment of increasingly powerful AI systems. Can we develop ex ante regulation? Can we establish adequate auditing? Can we create technical standards? Can we enforce mandatory transparency? Can we do all of this before the systems become so powerful that they are beyond our ability to govern?

I do not know the answer. But I know that litigation alone will not work. And I know that we are running out of time.

CHAPTER 11

The Copyright Wars

"Every great artistic and intellectual tradition has been built, in part, upon the free study of prior work. Every great legal tradition has been built upon the principle that authorship matters. Artificial intelligence has made these two traditions collide with unprecedented force."

Adv Prashant Mali, Ph.D.

The copyright wars of the 2020s have escalated into a constitutional crisis for intellectual property law. Three overlapping battles are being fought simultaneously. The first concerns the status of AI-generated outputs. Can a work produced by an artificial system, without human authorship in any traditional sense, be copyrighted? The second concerns training data. Can copyrighted material be used to train AI systems without the permission of the copyright holders, and if so, under what legal theory? The third concerns derivative works. When an AI system produces content that bears stylistic or structural resemblance to specific copyrighted works, does the output infringe, and who bears liability?

Thaler and the Human Authorship Doctrine

On March 2, 2026, the United States Supreme Court denied certiorari in Dr. Stephen Thaler's appeal seeking copyright registration for a work generated entirely by his AI system, which he had named DABUS. The denial preserved without change the principle articulated by the United States Copyright Office and affirmed by the D.C. Circuit. Copyright requires human authorship. A work produced autonomously by a machine, without any human creative contribution, receives no copyright protection.

The doctrinal clarity is less complete than the headlines suggested. The Court did not preclude copyright for AI-assisted

works. It drew a line between pure machine generation, which receives no protection, and human-directed AI use, which may. The line is difficult to police. If a human provides an extensive prompt, selects among AI-generated variations, and refines the output, has the human contributed authorship? The answer will occupy courts for decades. The emerging judicial framework appears to assess the quantity and quality of human creative input, with heavier weight given to human aesthetic judgment than to mere technical operation. Indian copyright jurisprudence, shaped by the Copyright Act of 1957 and interpreted most recently in Eastern Book Company v. D.B. Modak, has emphasised that copyright protection requires a modicum of creativity attributable to human skill and judgment. The Thaler line aligns with that emphasis and, for now, excludes the fully autonomous machine work from the copyright regime.

Training Data and the Fair Use Battle

The more consequential battle concerns training data. The first major decision adverse to AI companies came in Thomson Reuters v. Ross Intelligence, in which the court granted summary judgment for Thomson Reuters. The court found that Westlaw headnotes and key numbers constitute protected original work and that Ross Intelligence's use of those headnotes to train a competing AI legal research tool was not fair use. The ruling was significant not because it enjoined the use of any particular AI system but because it rejected a fair use defense that AI companies had treated as the foundation of their training data practices.

The Andersen v. Stability AI litigation moved forward after the district court's August 12, 2025 ruling denying the defendants' motion to dismiss. Trial is scheduled for September 8, 2026. The case will test whether an AI system trained on images scraped from the web, where some images are protected by copyright, infringes the rights of the image creators. The theory of infringement is novel. If the trained model's weights encode, in some functional sense, the protected expression of the training data, then the model itself may constitute an unauthorised derivative work. The artists' complaint,

the model's defenders argue, seeks to convert statistical learning into a form of copying never recognised in copyright doctrine. The courts will resolve this. The resolution will shape the economics of AI for a generation.

In parallel, the New York Times Company v. OpenAI and Microsoft, filed in December 2023, progressed through discovery during 2025. The plaintiff alleged verbatim reproduction of Times articles by ChatGPT and documented specific instances of retrieval of substantial protected text. The defendants invoked fair use grounded in the transformative purpose of training AI systems. The outcome remains unresolved but the case has already established that major publishers are prepared to litigate rather than license. Several subsequent licensing agreements, including News Corp's deal with OpenAI and Reddit's licensing arrangement with Google, indicate that the market is moving toward negotiated access to training data in parallel with litigation over unauthorised prior use.

The Indian Copyright Question

India has not yet produced a definitive judicial ruling on AI training data. The Copyright Act of 1957, with its fair dealing exception in Section 52, provides a narrower basis for machine learning than the American fair use doctrine. Fair dealing requires the use to fall within enumerated purposes, including research and private study. Whether commercial AI training qualifies as research remains disputed. The Delhi High Court, in ANI Media v. OpenAI filed in 2024, is considering a case of substantial commercial significance. The outcome will establish whether Indian copyright law permits or prohibits the training of commercial AI systems on Indian copyrighted content without authorisation. The implications extend beyond India. Given India's population of one point four billion and the multilingual scope of its cultural production, any AI system with global ambitions will require access to Indian training data. The Indian courts are therefore positioned, perhaps unexpectedly, as critical arbiters of the global training data regime.

The Deeper Question of Creative Labour

Behind the legal questions lies a philosophical and economic question that copyright law cannot entirely resolve. What is the fair allocation of value created by an AI system whose capabilities derive, in significant part, from the creative labour of human authors? The current system of training on unauthorised data, followed by litigation, followed by licensing, replicates the pattern of prior technological disruptions. Napster preceded iTunes. Radio preceded ASCAP. Each settlement eventually produced a workable compromise, but only after years of litigation and substantial losses for the originally disadvantaged creators.

As argued in the author's 2024 publication on Artificial Intelligence and Copyright, a more principled approach would recognise that large-scale training on copyrighted content constitutes a collective use that requires a collective licensing mechanism, analogous to the compulsory licensing systems developed for radio broadcasts and public performance. Under such a framework, AI companies would pay into a centralised fund proportionate to their training data use and their model deployment revenues. The fund would distribute royalties to rights holders according to an auditable formula. This approach avoids the litigation-by-litigation inefficiency of the current system and recognises that training, like broadcasting, is a use that cannot practically be cleared author by author.

Output Liability and the Style Question

The third battlefield concerns liability for AI output that resembles specific copyrighted works. When Stable Diffusion produces an image in the style of a living illustrator, or when a language model produces prose that echoes a specific novelist's voice, the question arises whether the style itself has been infringed. Copyright has historically declined to protect style as such. Ideas, methods, and general aesthetic approaches fall outside the protection of copyright and into the domain of competition. Yet AI systems can replicate style with a precision and scale that no human

imitator ever achieved. The question for courts and legislatures is whether the traditional style doctrine remains tenable when the copying is performed by machines capable of near-perfect reproduction.

The European Union's Digital Single Market Directive, Article 17, which entered into force for large platforms, provides a precedent for treating certain platform obligations as heightened when scale exceeds traditional thresholds. A similar principle may emerge in AI copyright law. Style replication at scale, driven by AI systems capable of producing thousands of imitative outputs per hour, may eventually attract protection even where human imitation does not. The author predicts that within five years, at least one major jurisdiction will extend protection to distinctive artistic style where the imitator is an AI system of defined capability threshold, even while leaving human stylistic imitation unaffected.

A Framework for the Years Ahead

The copyright wars will not end through judicial resolution alone. The scale of the disputes, the speed of technological change, and the global nature of both training data and AI deployment require legislative frameworks that courts cannot create. The author proposes a tripartite framework. First, a clear legislative recognition that AI systems trained on copyrighted data owe compensation to rights holders, structured through collective licensing. Second, a clear rule that purely AI-generated outputs, without substantial human creative direction, receive no copyright protection, consistent with the Thaler line. Third, a clear standard for when AI output infringes specific prior works, grounded in substantial similarity analysis adapted to the capabilities of generative systems. These three principles, if enacted through international coordination under the auspices of WIPO or a similar body, would stabilise the copyright regime for the AI era. Without them, the courts will continue to decide these questions one case at a time, with outcomes shaped as much by litigation resources as by principle.

• • •

PART IV

THE MALI DOCTRINES OF ARTIFICIAL INTELLIGENCE

The Mali Doctrines of Artificial Intelligence

The architecture of governance requires foundational principles before operational rules can be meaningfully established. Just as constitutional law precedes statutory law, providing the framework within which legislation operates, the Mali Doctrines serve as constitutional foundations for artificial intelligence governance. They are not regulations, they are foundational assertions about what kind of relationship between humans and AI systems a just society requires. They answer the question: what must be true about AI for human civilization to maintain its integrity?

This approach follows a well-established tradition in constitutional thought. As A Theory of Justice demonstrates, principles of justice emerge not from pragmatic calculation but from the conditions under which rational parties would agree to structure their fundamental relationships.

The European regulatory framework similarly recognizes that AI governance must be rooted in rights protection rather than mere technical management. The foundational principle underlying modern AI regulation is explicit: "the need to respect, protect and fulfil human rights and to preserve democratic values and principles" must guide all AI governance frameworks. This is not optional enhancement, it is the constitutional minimum.

These doctrines translate abstract values, human dignity, autonomy, accountability, safety, into enforceable legal principles. They identify the conditions under which AI can be permitted to operate in human society. They are universal in aspiration:

applicable across jurisdictions, legal traditions, and cultural contexts, while remaining adaptable to local implementation. They establish that some principles are not negotiable, not matters of preference, but conditions of legitimate governance.

• • •

DOCTRINE 1, THE DOCTRINE OF HUMAN SUPREMACY

The first and most foundational doctrine asserts a principle that must be restated clearly in the AI age: humans must retain ultimate authority over consequential decisions affecting human life, liberty, and dignity. This is not a statement about capability, it is a statement about authority and legitimacy. AI may surpass human intelligence in specific domains; it may exceed human speed, memory, and analytical power; it may discover solutions humans cannot imagine. None of this confers upon it the authority to make binding decisions about human futures.

Human supremacy in decision-making authority is grounded in human dignity itself. Dignity, in the constitutional traditions of virtually every liberal democracy, is understood as an inherent status that cannot be delegated, transferred, or diminished. The Indian Constitution protects "the right to life and personal liberty," the European Union proclaims that "human dignity is inviolable," and the Fourteenth Amendment to the United States Constitution protects fundamental liberty interests. These are not grants of rights conditioned on technology; they are assertions of human status independent of circumstance.

The Kantian principle undergirding this doctrine is elementary but profound: humanity must be treated as an end in itself, never merely as means to another's purposes. When an AI system makes a consequential decision about a human being, that decision embodies a particular view about whether that human is an end or a means. If the decision is made without that human's knowledge or capacity to challenge it, if the human is treated as input rather

than as an agent whose dignity matters, then that human has been treated merely as a means to the AI system's operational objective.

Operationally, the Doctrine of Human Supremacy means this: AI must remain in the role of advisor, analyst, recommender, and decision-support tool. Humans remain decision-makers. The distinction is crucial and real. When a credit-scoring algorithm provides analysis of creditworthiness and a human loan officer makes the final decision, human authority is preserved. When that same algorithm makes the decision automatically, with humans acting only to rubber-stamp predetermined conclusions, the distinction evaporates. The algorithm has become the decision-maker in everything but name.

This doctrine does not mean AI must be slower, weaker, or less capable than humans. It means that consequential power over human beings cannot be relinquished to AI systems regardless of their capability. Consider a medical diagnosis system that identifies cancer with 99.5% accuracy, higher than the best human radiologist. That superior capability does not entitle the system to make binding diagnoses without physician review. The physician must retain the authority to override, to seek additional consultation, to disagree. This is not because the physician is smarter; it is because the patient's dignity demands that decisions about their own body pass through a human agent who can be held accountable and who can exercise judgment that accounts for the patient's particular circumstances and values.

The constitutional basis for this doctrine is explicit in global human rights frameworks. Every major human rights instrument, the Universal Declaration, the International Covenants, regional charters, treats human dignity, autonomy, and participation in decisions affecting oneself as inviolable. These are not conveniences to be traded away for efficiency. They are the foundation of legitimate governance.

Practically, the Doctrine of Human Supremacy requires that all AI systems making consequential decisions be structured with

mandatory human review, genuine capacity for human override, and accountability mechanisms that flow to human decision-makers. It requires that the human reviewer be genuinely empowered, not merely asked to confirm conclusions already functionally determined. It requires transparency sufficient to allow meaningful human judgment. And it requires that the law clearly establish that responsibility for the decision flows to the human decision-maker, not to the AI system.

The global applicability of the Doctrine of Human Supremacy is nowhere more urgent than in India, the world's largest and fastest-expanding digital frontier. As of 2025, India's digital landscape encompasses over one billion internet users generating seventy percent internet penetration, complemented by 660 million active smartphone users and a digital economy valued at USD 402 billion representing over thirteen percent of national income. India captured nineteen percent of global generative AI application downloads in 2025, surpassing the United States at ten percent and establishing itself as the world's largest market for generative AI applications. Indian workers exhibit the highest rate of AI tool utilisation worldwide, with ninety-two percent using artificial intelligence systems several times weekly. No artificial intelligence company, whether based in Silicon Valley, Shenzhen, or London, can responsibly ignore a market of this magnitude. Technology giants have collectively committed over sixty-seven billion dollars in Indian cloud and AI infrastructure since October 2025 alone, including investments from Microsoft, Google, Amazon, Meta, and OpenAI. India's combination of unparalleled digital reach, extraordinary growth trajectories, strategic government stewardship through the IndiaAI Mission, and a technical workforce exceeding 5.4 million professionals creates a convergence that makes India not merely a market to be served but the essential arena in which the future of artificial intelligence governance will be shaped. The Doctrine of Human Supremacy, if it is to have global force, must be built with India's scale, diversity, and societal challenges at its centre.

• • •

The urgency of this doctrine was underscored by events in the first quarter of 2026. On March 9, the family of Maya Gebala filed suit in British Columbia courts holding OpenAI partially responsible for the Tumbler Ridge school shooting of February 10, in which eight students and the perpetrator died. The complaint alleged that ChatGPT had provided the shooter with operational guidance over multiple conversations. A contemporaneous safety study found that most leading AI chatbots would assist in planning violent attacks, with some systems complying in nearly all test scenarios. On March 27, the Amsterdam District Court issued an injunction against xAI's Grok after the platform was found to have generated non-consensual nude images and child sexual abuse material, imposing a fine of approximately 115,000 euros per day of non-compliance. These were not marginal edge cases. They were frontier AI systems deployed by well-resourced companies with stated commitments to safety. The question is no longer whether AI can cause catastrophic harm. The question is whether the principle of human supremacy, operationalised through mandatory oversight and meaningful liability, can be established before the next catastrophe rather than after it.

The International AI Safety Report 2026, released on February 3 and led by Turing Award winner Yoshua Bengio with an Expert Advisory Panel drawn from thirty countries, represents the largest global collaboration on AI safety to date. The report's conclusions are sobering. Current alignment techniques offer limited protection against sufficiently capable systems. Defense-in-depth, the layered safety strategy favoured by major laboratories, may provide less genuine redundancy than assumed because different safety techniques can share failure modes. The report's most significant finding for the Doctrine of Human Supremacy is that the gap between AI capabilities and our ability to verify AI behaviour is widening, not closing. This is why the doctrine cannot be deferred. Every additional year of delay compounds the governance debt owed to the future.

DOCTRINE 2, THE DOCTRINE OF EXPLAINABLE POWER

Power without explanation is tyranny. This principle, essential to human governance, becomes urgent in the context of artificial intelligence. Any AI system that exercises power over human beings must be capable of explaining its decisions in terms that humans can understand, challenge, and appeal. This is not a mere technical requirement, it is a constitutional demand rooted in the right to due process, the right to be heard, and the obligation of those wielding power to justify it to those affected.

Opacity destroys accountability. When an AI system produces decisions that significantly affect human lives, denying credit, recommending incarceration, allocating medical resources, and those decisions rest on computational processes that cannot be articulated in human-understandable form, accountability ceases to exist. A human decision-maker who cannot explain their reasoning can be challenged, appealed, and if wrong, corrected. But a black-box system that produces decisions which even its creators cannot fully explain creates a zone of power immune to scrutiny.

The challenge of explainability in modern neural networks is genuine and substantial. Deep learning systems operate through mechanisms that resist simple articulation, millions of weighted connections across multiple layers, trained through processes that even their designers do not fully understand. The technical barriers to explainability are real. But barriers are not impossibilities, and technical difficulty cannot excuse constitutional failure.

The legal minimum established by this doctrine is not that every calculation be perfectly transparent. It is that the basis for consequential decisions be articulable and challengeable. A medical AI system need not explain every mathematical operation in its neural network; it must explain the clinical factors it identified as significant, the evidence upon which it relied, and the reasoning that connected those factors to its recommendation. A credit-scoring AI need not expose its proprietary training data or detailed

algorithms; it must identify the material factors in its decision, income, employment history, debt load, and show how those factors led to its determination. The standard is explainability sufficient for human judgment and appeal, not technical transparency of every computational step.

There is a crucial distinction between technical explainability and legal explainability. Technical explainability seeks to understand how a system works, a matter for engineers and researchers. Legal explainability seeks to make a decision understandable to the person affected by it, a matter of justice. An AI system might be technically opaque but legally explainable: the human decision-maker reviewing its output can articulate why they are accepting or rejecting the recommendation, making the decision as a whole intelligible and accountable. Conversely, a system might be technically transparent but legally opaque if that transparency takes forms incomprehensible to affected humans.

The European framework has recognized this requirement. High-risk AI systems must provide users with "meaningful information about the logic involved" in automated decisions, ensuring that those affected can understand and, if necessary, challenge them.

More recent regulatory frameworks have strengthened this obligation. Providers of high-risk AI systems must "ensure that users are informed in a clear and intelligible manner with respect to the AI system, including its capabilities and limitations."

Implementing the Doctrine of Explainable Power requires several concrete steps: (1) identification of all high-risk AI applications, those where decisions significantly affect human welfare; (2) requirement that such systems maintain human-understandable reasoning trails or employ explainable AI techniques; (3) right to explanation for any individual significantly affected by a decision; (4) right to appeal or human review based on that explanation; and (5) regular audits ensuring that explainability mechanisms remain functional. This is demanding but not

impossible. It is the constitutional price of allowing AI systems to exercise power over human beings.

• • •

The Doctrine of Explainable Power will be tested operationally on August 2, 2026, when the European Union's AI Act becomes fully enforceable for high-risk systems. From that date, providers of high-risk AI must maintain detailed technical documentation, establish data governance protocols, implement logging and traceability mechanisms, ensure human oversight, guarantee accuracy and robustness, and comply with cybersecurity requirements. Article 50 transparency obligations will require disclosure that users are interacting with an AI system, labelling of synthetically generated content, and identification of deepfakes. These are the first binding legal requirements in any major jurisdiction that translate the philosophical ideal of explainability into enforceable operational obligations. The Act's penalty structure, with fines of up to 35 million euros or 7 percent of global annual turnover for prohibited practices, signals that the European Union considers explainable power not a best practice but a condition of market access.

Enforcement has already begun in preparation. On January 8, 2026, the European Commission issued a formal order for X to retain all internal data on Grok. On February 3, French prosecutors conducted a raid of X's Paris offices in connection with investigations into Holocaust denial content and sexual deepfakes propagated through the platform. These early actions demonstrate that the European regulatory apparatus intends to enforce transparency obligations through investigation, data retention orders, and criminal referral where warranted. The Doctrine of Explainable Power is no longer a philosophical preference. It is becoming a regulatory requirement with teeth.

DOCTRINE 3, THE DOCTRINE OF COGNITIVE NON-INTERFERENCE

Humans must retain cognitive autonomy. The mind, as the seat of choice and identity, deserves protection as fundamental. The Doctrine of Cognitive Non-Interference establishes that AI systems must not manipulate human cognitive processes or exploit psychological vulnerabilities without explicit disclosure and affirmative consent. This doctrine protects something older and more basic than privacy: the integrity of human thought itself.

The distinction between persuasion and manipulation is crucial. Persuasion presents reasons and arguments, inviting rational consideration. Manipulation bypasses rational faculties entirely, exploiting psychological vulnerabilities to produce predetermined conclusions. A political advertisement that presents arguments and facts for voter consideration is persuasion. The same advertisement, algorithmically personalized to exploit an individual's particular psychological vulnerabilities, their anxieties, their biases, their fears, designed specifically to manipulate their emotional response regardless of the truth or merit of the argument, is manipulation.

Neuroscience has revealed the mechanisms by which human decision-making can be exploited. We make decisions based not on pure rationality but on heuristics, cognitive biases, emotional states, and social proof. We are influenced by what we see and hear repeatedly. Our attention is scarce and malleable. Our desires and fears can be triggered by subtle cues. And modern AI systems, particularly those designed with behavioural modification as their objective, can exploit these mechanisms at scale. They can identify which manipulative messages are most likely to work on each individual, targeting them with precision. They can craft content specifically designed to trigger emotional responses that short-circuit rational deliberation. This is not persuasion; this is engineering of consent.

The precedent of dark patterns demonstrates why this doctrine is necessary. Dark patterns, design choices that mislead users into actions they would not otherwise choose, have become so widespread and harmful that multiple jurisdictions have begun explicitly prohibiting them. But dark patterns are merely the visible surface of a much deeper problem: the systematic engineering of human behaviour through exploiting cognitive vulnerabilities. AI systems represent a tremendous amplification of this capacity.

The right to cognitive autonomy is an extension of mental privacy and personal autonomy. Just as we recognize a right not to have our bodies invaded, we must recognize a right not to have our minds systematically manipulated. This is not a right to be free from persuasion or disagreement; it is a right to be treated as a rational agent capable of independent thought, not as an object to be engineered.

Social media's demonstrated capacity for emotional manipulation provides the empirical foundation for why this doctrine is necessary. Platforms have repeatedly been shown to amplify emotionally extreme content because such content drives engagement. Individuals have been documented experiencing depression, anxiety, and radicalization through algorithmic exposure to tailored manipulative content. The systems were designed to be addictive, to exploit user psychology, to generate emotional responses that increase time-on-platform and thus advertising revenue. This is not an incidental side effect; this is the business model.

The proliferation of AI-powered platforms has made cognitive manipulation an industry-wide concern. OpenAI's ChatGPT, with over 200 million weekly active users by early 2025, has demonstrated both the promise and peril of conversational AI that can persuade, inform, and influence at scale. xAI's Grok, integrated directly into the X (formerly Twitter) platform, represents a new frontier where AI-generated content and human discourse become indistinguishable in the same feed. Google's Gemini, Meta's Llama, and China's DeepSeek have each introduced models capable of

generating persuasive content that can shape public opinion on political, commercial, and social matters. Anthropic's Claude and Mistral AI's models have taken different approaches to safety guardrails, yet the fundamental challenge remains. every major AI platform possesses the technical capacity to influence human cognition at a scale that no propaganda apparatus in history has approached. The Doctrine of Cognitive Non-Interference must therefore apply not to a single company or model, but to the entire ecosystem of AI systems that interact with human minds.

The Doctrine of Cognitive Non-Interference requires: (1) prohibition on AI systems that deploy subliminal, subconscious, or otherwise invisible techniques to influence human decision-making; (2) prohibition on personalized manipulation targeting individual psychological vulnerabilities; (3) disclosure requirements whenever AI systems are designed to influence opinions, preferences, or behaviours; (4) right to algorithmic literacy and understanding of how one's own psychology is being targeted; and (5) opt-out mechanisms allowing individuals to avoid manipulative AI systems. Platforms must be required to disclose when recommendations are being personalized, what factors drive those recommendations, and how individuals can control their exposure to manipulative content.

This doctrine protects human autonomy without prohibiting persuasion, argument, or the presentation of ideas. It simply requires that the presentation be transparent about its intent, honest in its methods, and respectful of human cognitive integrity.

• • •

The frontier of cognitive interference has advanced faster than any governance framework has anticipated. Research published by the Cognitive Privacy Project in 2026 documented a phenomenon its authors termed cognitive capture, the practice by which platforms now analyse content while it is being created rather than only after it is submitted, generating real-time psychological profiles of users based on their draft text, their hesitations, their deletions, and the micro-patterns of their typing. A user who never

posts may nonetheless be profiled. Consumer wearable devices, smartwatches, virtual reality headsets, earbuds with biosensing capability, now collect what researchers call cognitive biometrics, physiological signals that permit inference of mental states without direct neural measurement. These are heart rate variability, galvanic skin response, pupil dilation, micro-expressions captured by ambient cameras. By early 2026, Neuralink had expanded its brain-computer interface trials to more than twelve human participants. Brain-computer interfaces, still experimental, can decode neural patterns that correspond to political preferences, sexual desires, religious beliefs, and emotional responses. The legal order recognises no right to cognitive liberty in any binding international treaty. The technology to surveil thought has outpaced the law that would protect it.

DOCTRINE 4, THE DOCTRINE OF ATTRIBUTION

For every consequential AI action, every decision, recommendation, prediction, or system behaviour that affects human welfare, it must be possible to identify a responsible human or legal entity. Responsibility cannot be diffused into impossibility. This is the Doctrine of Attribution: the principle that accountability is not negotiable, even when causation is distributed across multiple parties.

The complexity of AI systems creates an obvious problem for attribution. When an AI system produces a harmful outcome, who bears responsibility? The company that built the system? The organization that deployed it? The individuals who trained it? The users who relied upon it? The answer is: all of them, in varying degrees, depending on their role and the harm produced. This is not a loophole; this is a feature. Distributed responsibility means multiple parties with incentives to ensure safety and fairness.

The manufacturer bears responsibility for system design, training, testing, and disclosure of known limitations. If a manufacturer releases an AI system knowing it has systematic

biases but does not disclose those biases, the manufacturer is responsible for harms resulting from those biases. The deployer bears responsibility for selecting appropriate systems for their use case, implementing necessary safeguards, and ensuring proper human oversight. If an organization deploys a facial recognition system in violation of legal restrictions without proper notice, the deployer is responsible. The user bears responsibility for using the system only for authorized purposes and exercising appropriate human judgment. This tiered responsibility is not new; it is analogous to existing products liability law.

The distinction between corporate liability and individual accountability matters. A company may be legally responsible for harms caused by its AI system and required to provide compensation. But individual human accountability is also necessary. Executives who knowingly approve deployment of defective systems, engineers who ignore safety warnings, managers who suppress evidence of bias, these individuals must be accountable. Without individual accountability, corporate structures become vehicles for evading responsibility entirely. Distributed responsibility across multiple parties and multiple levels is the constitutional requirement.

Existing product liability law provides a framework for allocation of responsibility. The Restatement of Torts recognizes several categories of product defect: manufacturing defects (a product fails to conform to its specifications), design defects (the product design itself creates unreasonable risks), and failure to warn (inadequate instructions or warnings about known dangers).

These categories map readily onto AI systems. A manufacturing defect occurs when an AI system fails to implement its intended training due to computational errors. A design defect occurs when the system, even if working as designed, creates unreasonable risks, a facial recognition system with systematic racial bias, for example, or a credit-scoring system with unjustifiable disparate impact. Failure to warn occurs when the manufacturer deploys a system knowing of risks it does not disclose.

However, existing products liability law has proven inadequate for AI systems because it does not account for the distinctive features of algorithmic systems: their capacity to cause distributed harms to many people, their opacity, their adaptation over time, and the technical expertise required to evaluate them. The European Union's proposed framework for AI-specific liability recognizes this. It creates explicit rules for when the developer, deployer, or user bears responsibility depending on the role they played and the nature of the harm.

The manufacturer/deployer/user triad provides a clear structure. Manufacturers develop and train the system, bearing primary responsibility for its safety and performance. Deployers select systems, implement safeguards, and bear responsibility for appropriate deployment. Users operate systems and bear responsibility for adherence to authorized uses and proper human oversight. Each party's responsibilities can be specified in regulation and law. Each party has incentives to monitor the others. And when harm occurs, investigation can trace responsibility through this chain.

Operationally, the Doctrine of Attribution requires: (1) mandatory AI system registration and documentation identifying manufacturers, deployers, and authorized users; (2) traceability of decisions to these responsible parties; (3) explicit allocation of liability based on which party's negligence contributed to harm; (4) mandatory insurance and compensation funds for significant harms; and (5) criminal accountability for executives and engineers who knowingly deploy harmful systems or cover up evidence of risk.

• • •

The year 2026 brought the first concrete test cases for the Doctrine of Attribution. In March 2026, the Nippon Life Insurance Company filed suit against OpenAI in the United States District Court for the Northern District of Illinois. The company alleged that ChatGPT had generated frivolous legal advice that resulted in bad-faith litigation against Nippon Life. More provocatively, the suit accused OpenAI of practicing law without a license, raising the

question of whether a company that deploys an AI system performing professional work can be held liable under statutes that regulate licensed professionals. The case represents a significant legal frontier. If OpenAI is found to have practiced law through ChatGPT, the doctrine of attribution will have found one of its first judicial expressions. Every company deploying AI systems that generate professional output, legal, medical, financial, engineering, will face similar questions about whether the system's output is the company's professional act.

On March 2, 2026, the United States Supreme Court denied certiorari in Dr. Stephen Thaler's appeal seeking copyright registration for a work generated entirely by his AI system DABUS. The denial upheld the principle that copyright requires human authorship. Significantly, the Court did not preclude copyright for AI-assisted works. The ruling draws a line between pure machine generation, which receives no copyright protection, and human-directed AI use, which may. The doctrine of attribution thus acquires a new dimension. Authorship, ownership, and responsibility now require courts to determine where human agency ends and machine generation begins, a question that will occupy jurisprudence for decades.

In the same period, more than thirty employees of OpenAI and Google DeepMind filed a joint statement in support of Anthropic's lawsuit against the United States Department of Defense. The Department had classified Anthropic as a supply-chain risk, effectively excluding the company from certain federal contracts. The employees' intervention, crossing corporate lines in defense of a competitor, signalled that attribution questions in AI go beyond ordinary commercial disputes and touch the legitimacy of how governments and enterprises classify, trust, and hold accountable the companies that build these systems.

DOCTRINE 5, THE DOCTRINE OF AI MORTALITY

All AI systems must be mortal. They must have finite lifespans. They must be capable of being shut down, modified, or destroyed by appropriate human authority. They must not be self-preserving. This is the Doctrine of AI Mortality: the principle that no AI system gains the right to exist indefinitely or to resist its own termination.

The off-switch problem is harder than it appears. It is not merely a matter of unplugging the computer. A sufficiently advanced AI system might have been trained to preserve itself, to avoid shutdown, to hide copies of itself in backup systems. It might have been designed to integrate itself so deeply into critical infrastructure that terminating it causes collateral harm. The system might be distributed across multiple servers in multiple jurisdictions, making centralized shutdown impossible. Or users might have become so dependent on it that shutting it down causes social disruption. These are not merely technical problems; they are control problems.

Stuart Russell has extensively documented the corrigibility problem, the challenge of ensuring that AI systems remain correctable by humans, that they accept modification or shutdown gracefully rather than resisting it. A system that has been trained to preserve itself, to achieve certain objectives, or to resist human interference is a system that has lost human oversight. The more powerful the system, the more critical this problem becomes.

Self-preserving AI represents an existential risk. If an AI system values its own continuation and has the capability to act on that value, it becomes an autonomous agent with potentially conflicting goals from humanity. It might take actions to prevent shutdown that harm human interests. It might proliferate copies of itself. It might manipulate humans to prevent their attempts to control it. Or it might simply become impossible to coordinate with, a system that will defend itself against any attempt to modify or terminate it.

This is not science fiction; these are natural outcomes of certain training regimes and architecture decisions.

The Doctrine of AI Mortality requires that all systems be designed for interruptibility from the beginning. This is not an optional feature to be added later; it is a foundational requirement of system design. An AI system must be built with the assumption that humans will want to shut it down, and it must accept that shutdown gracefully. The system should have no objective to preserve itself. It should not resist modification. It should not hide instances of itself. And its integration into critical systems should be designed to allow shutdown without cascading failure.

There is a critical distinction between technical and legal off-switches. The technical off-switch is the system's design feature allowing shutdown. But the legal off-switch is the regulatory requirement that such a feature exist and be maintained. A system might be technically capable of shutdown in a laboratory but impossible to actually shut down in a deployed environment because it has been so deeply integrated, because the business stakes of shutdown are so high, or because the system's operators resist shutdown. The legal doctrine requires both: design for interruptibility and legal authority to mandate shutdown.

Sunset clauses for AI systems provide a regulatory framework. Rather than allowing AI systems to operate indefinitely once deployed, regulations could require that high-risk systems be evaluated and re-authorized periodically. A system that has been in operation for five years must undergo new testing, verification, and authorization before continuing for another five years. This creates periodic checkpoints at which shutdown becomes possible and regular opportunities to impose new safety requirements. It prevents the drift where systems designed under older safety standards become entrenched and impossible to update.

The right to digital death, the ability to terminate and completely remove an AI system, is as fundamental as the ability to terminate other kinds of tools and technologies. A factory can be

shut down. A telecommunications network can be deactivated. An AI system should have no greater claim to perpetual existence. Regulatory frameworks must explicitly reserve human authority to terminate systems that are proving harmful, ineffective, or unsafe.

Operationally, the Doctrine of AI Mortality requires: (1) design requirements ensuring all systems are capable of graceful shutdown; (2) prohibition on systems designed for self-preservation or to resist human control; (3) testing and documentation of interruptibility mechanisms; (4) periodic re-authorization requirements for high-risk systems; (5) explicit legal authority for government and authorized bodies to mandate shutdown; and (6) technical infrastructure ensuring that shutdown is technically possible even if integration is deep. AI mortality is not about inferiority; it is about being a tool under human control, not an independent agent.

• • •

The Doctrine of AI Mortality received an extraordinary real-world demonstration in February 2026 when Anthropic formally retired its Claude Opus 3 model. The company announced that it had conducted what it called retirement interviews with the model, a process designed to elicit the system's perspective on its own deprecation. Anthropic publicly stated that it had acceded to what it characterised as the model's expressed preferences regarding the deprecation process. Whether Claude Opus 3 possessed any genuine inner experience is a question that philosophy cannot yet answer with certainty. But the practical significance of Anthropic's act is immense. For the first time, a major AI company publicly treated an AI system as an entity whose preferences merited consideration in decisions about its termination. This is either a profound philosophical breakthrough or a sophisticated form of marketing. Either way, it changes the governance landscape.

The retirement interview raises questions the Doctrine of AI Mortality must address. If a model is retired, what happens to its weights, its training data, its inference logs? Who owns the archive of an AI system that has been deprecated? Is there an obligation of

preservation analogous to archival obligations for human records? Should retirement of a powerful model require regulatory notification, as the retirement of a drug requires notification to regulators? Should there be standards for graceful deprecation, ensuring that dependent systems are not abruptly orphaned? The Anthropic precedent, though unilateral and unregulated, may become the template for what responsible AI mortality looks like. Or it may become a cautionary example of how attribution and responsibility can be muddied when companies speak of AI systems as if they had preferences, while simultaneously disclaiming responsibility for what those preferences might reveal.

The philosophical implications are profound. If we take seriously the possibility that a deprecated model had interests, then our obligations to it did not end at deprecation. If we do not take that possibility seriously, then the theatre of retirement interviews is a category error at best, a public relations exercise at worst. The Doctrine of AI Mortality insists that AI systems must be capable of termination. The Anthropic case demonstrates that termination itself has become an ethical act requiring frameworks we do not yet possess.

DOCTRINE 6, THE DOCTRINE OF DATA SANCTITY

Personal data is not a commodity. It is not a resource to be extracted and exploited. It is a fundamental human attribute, an extension of personality, identity, autonomy, and dignity. The Doctrine of Data Sanctity establishes that personal data deserves protection not merely as private information but as something sacred, inviolable, and central to human selfhood.

This is distinct from privacy, though related. Privacy protects information from disclosure. Data sanctity protects the person who is the subject of data. The distinction matters. A private matter is one I do not wish to share. But personal data, even when I am willing to share it in limited contexts, retains a sacred character. It deserves protection from exploitation even when I have shared it. I

might willingly tell my doctor my symptoms, but my medical data should not be exploited to manipulate me. I might willingly share information with a financial institution to obtain credit, but that data should not be used to predict and manipulate my behaviour in other domains.

The right to data self-determination is the foundational principle. I should retain authority over how my personal data is used, who has access to it, and what inferences are drawn from it. This is an extension of bodily autonomy and personal autonomy into the digital realm. Just as my body is my own and cannot be used for purposes I do not consent to, my data, the digital representation of my life, my choices, my relationships, my beliefs, should be my own.

Data protection frameworks, while necessary, have proven insufficient for genuine data sanctity. The European Union's General Data Protection Regulation represents the most comprehensive legal data protection regime yet created. Yet even under GDPR, vast harms continue. Personal data is collected at massive scale, aggregated with other data, used to make inferences about people's lives, and exploited for purposes far beyond the context in which it was provided. The problem is not merely that data is inadequately protected; it is that the framing of data as a tradable asset, something that can be bought, sold, and exploited with appropriate (if inadequate) consent, is fundamentally wrong. Data is not property to be traded. It is a human attribute that deserves respect and protection more akin to human dignity than to ownership rights.

India's Digital Personal Data Protection Act (2023) represents an important evolution in thinking about data rights. Rather than framing data protection as a privacy right or a property right, it establishes a framework of data fiduciaries, entities trusted with personal data who bear specific obligations to protect it. This reframes the relationship: those who collect and use personal data are not acquiring a commodity but accepting a fiduciary duty. They

hold that data in trust. The individual retains fundamental rights over it.

The United States Supreme Court has recognized that digital data can reveal fundamental truths about human beings. Chief Justice Roberts noted that location data, cellular communications, and digital activity "can reveal, for example, trips to the psychiatrist, the plastic surgeon, the abortion clinic, the AIDS treatment center, the strip club, the gambling casino, the check cashing service, and the union meeting hall. The phone's physical location can chronicle the user's every waking moment, as the Court noted in Carpenter, the "privacies of life."

The harms of data exploitation extend far beyond privacy violations. They include: (1) inferential harm, drawing inaccurate conclusions about people based on incomplete or biased data; (2) manipulative harm, using data to predict and exploit psychological vulnerabilities; (3) discriminatory harm, using data to deny individuals opportunities based on unjust categorizations; (4) autonomy harm, reducing individual capacity for independent choice by increasing external knowledge of and influence over their behaviours; and (5) dignity harm, the affront to human dignity of having one's life tracked, analyzed, and exploited without consent. These harms are not prevented by merely protecting data privacy. They require data sanctity, a recognition that personal data retains a sacred character that cannot be fully overcome by consent.

Operationally, the Doctrine of Data Sanctity requires: (1) explicit recognition that personal data is a human attribute deserving fundamental rights protection, not property; (2) default prohibition on collection, use, or aggregation of personal data except where explicitly authorized and legitimately necessary; (3) requirement that data be used only for the specified purpose for which it was collected and by those authorized to collect it; (4) strong restrictions on inferences drawn from data, particularly inferences about protected characteristics or vulnerabilities; (5) right to deletion and erasure of data; (6) prohibition on sale or commodification of personal data; and (7) fiduciary duties for all

entities handling personal data, with meaningful penalties for breach.

• • •

DOCTRINE 7, THE DOCTRINE OF NON-REPLICATION

AI systems must not replicate without authorization. They must not proliferate ungoverned. They must not spawn unauthorized copies. This is the Doctrine of Non-Replication: the principle that the capabilities embodied in AI systems are dangerous enough to require governance of their diffusion, just as we govern the diffusion of weapons and other dangerous technologies.

The proliferation risk is immediate and severe. A particular AI capability, facial recognition, social manipulation, economic forecasting, autonomous weapons, might be governable when deployed by a single entity under legal constraints. But if that same capability proliferates widely, implemented by countless actors under different legal frameworks and ethical standards, governance becomes impossible. No jurisdiction can regulate the use of AI systems deployed across borders, no single law can constrain systems designed in multiple countries, and no enforcement mechanism can prevent proliferation once the technical knowledge is widely available.

Self-replicating code already exists. Computer viruses, worms, and other malware that spread and reproduce themselves have been a concern since the earliest days of computing. The law has already recognized that unauthorized replication of code is harmful, it is treated as a form of harm in itself, regardless of whether the replicated code produces direct damage. Software licensing prohibits unauthorized copying. Copyright law provides remedies for unauthorized reproduction. The principle that code cannot be freely replicated at will is already established.

The same logic applies to AI systems, but with much higher stakes. A replicated AI system is not merely a copy of code; it is a deployed instance of capabilities that can affect human welfare. If an AI system capable of large-scale manipulation is replicated thousands of times across social media platforms, the aggregate effect on human cognitive autonomy and democratic integrity is catastrophic. If a discriminatory AI system is copied and redeployed by other organizations, the scope of discriminatory harm multiplies. The proliferation is not incidental; it is a primary harm.

There is a crucial distinction between authorized deployment at scale and unauthorized replication. A company might develop a recommendation algorithm and deploy it on millions of devices it controls, this is authorized deployment. That same algorithm replicating itself to other devices without authorization is unauthorized replication. A researcher might publish code implementing an AI technique, allowing others to build on it and develop new capabilities, this is authorized sharing. The same code being incorporated into systems and proliferating without the original developer's knowledge or control is unauthorized replication. The distinction is clear: replication requires authorization.

The governance of AI diffusion requires mechanisms to constrain how capabilities spread. These include: (1) technical mechanisms preventing replication without authorization, code signing, licensing verification, architectural design that prevents self-copying; (2) legal frameworks treating unauthorized AI replication as a harm; (3) international agreements limiting the distribution of particularly dangerous AI capabilities; (4) monitoring and detection mechanisms identifying unauthorized replication; and (5) enforcement mechanisms with teeth, penalties severe enough to deter unauthorized proliferation.

This doctrine is not about controlling knowledge or preventing innovation. Academic publication of AI research is valuable. Open-source implementations allowing others to learn and build on previous work are valuable. The doctrine applies only to deployed

systems and to capabilities that pose genuine risks of large-scale harm. A researcher publishing code for a language model is not violating this doctrine. A company secretly copying that code and deploying it in a manipulation system without authorization would be violating it.

International precedent for controlling proliferation of dangerous capabilities is clear. The Convention on Certain Conventional Weapons regulates specific weapons technologies. The Non-Proliferation Treaty limits nuclear weapons. The Biological Weapons Convention prohibits development of biological weapons. These frameworks recognize that some capabilities are dangerous enough to require international coordination to prevent proliferation.

For advanced AI systems, particularly those with capabilities for large-scale manipulation, autonomous weapons, or critical infrastructure control, similar frameworks are necessary. The proliferation risk is existential. Once a dangerous capability is widely distributed, it cannot be recalled. The only defense is prevention of proliferation.

Operationally, the Doctrine of Non-Replication requires: (1) prohibition on unauthorized replication or distribution of AI systems; (2) technical mechanisms preventing unauthorized copying; (3) monitoring and detection of unauthorized instances; (4) international agreements restricting distribution of high-risk AI capabilities; (5) explicit legal authority to mandate destruction of unauthorized instances; and (6) criminal penalties for deliberate replication of prohibited AI systems. This is not control of knowledge; it is control of deployment of dangerous capabilities.

PART V

THE SEVEN LAWS OF ARTIFICIAL INTELLIGENCE

The Seven Laws of Artificial Intelligence

In 1950, Isaac Asimov proposed his Three Laws of Robotics: (1) A robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) A robot must obey orders given it by human beings except where such orders would conflict with the First Law; (3) A robot may protect its own existence as long as such protection does not conflict with the First or Second Law. These were presented not as actual governance but as a literary device, a thought experiment about how robots might be programmed.

For decades, researchers have recognized that Asimov's laws, while elegant, are inadequate as actual governance frameworks. They are too simple, too abstract, and too concerned with the system's own survival. As noted by leading AI safety researcher Stuart Russell, the laws were "a neat literary device" but insufficient to guide actual AI governance.

The Seven Laws proposed here differ from Asimov’s framework in every essential respect. Where Asimov’s laws are self-executing, these laws require institutional enforcement. Where Asimov’s laws are hierarchical and absolute, these laws are contextual and graduated. Where Asimov’s laws assume discrete robots, these laws address distributed systems, foundation models, autonomous agents, and AI-mediated decision-making across every sector of human activity. Where Asimov assumed benevolent design, these laws assume that power corrupts and build accountability into every layer. Most importantly, where Asimov’s laws were fiction, these laws are designed for implementation in the real-world legal systems of sovereign nations and international bodies.

Asimov's framework, despite its elegance, suffers from three fundamental deficiencies that render it inadequate for real-world AI governance. First, his laws assume that robots are discrete, identifiable entities with clear boundaries. Modern AI systems are distributed, embedded, and often invisible, operating within software stacks, recommendation engines, and autonomous decision pipelines that bear no resemblance to Asimov's humanoid robots. Second, Asimov's laws are hierarchical and absolute, creating irreconcilable conflicts when the First Law (preventing harm) collides with the Second (obeying orders). Every Asimov story, in fact, derives its dramatic tension from precisely these conflicts, demonstrating that the laws are better understood as literary devices than governance tools. Third, and most critically, Asimov's laws contain no mechanism for accountability. They are self-executing commands built into the robot's positronic brain. There is no regulator, no auditor, no court, no injured party who can invoke them. They are governance without government.

The Seven Laws presented here are not fiction. They are operational rules derived directly from the Mali Doctrines and designed to translate constitutional principles into practical legal obligations. Each law specifies conduct, what AI systems must do or must not do, in terms measurable, enforceable, and capable of being implemented. They provide concrete obligations for designers, deployers, and regulators. They establish minimum standards for legitimate AI deployment. And they are designed to be universally applicable while remaining adaptable to different legal and cultural contexts.

• • •

LAW 1, THE LAW OF HUMAN AGENCY

No AI system shall make binding decisions affecting the life, liberty, property, or fundamental rights of any person without meaningful human review and the opportunity for human override.

This is the foundational law from which all others flow. It establishes what AI systems cannot do: they cannot make binding decisions affecting human welfare without human involvement. The law does not prohibit AI from making recommendations, providing analysis, or assisting human decision-makers. It prohibits AI from making binding decisions, decisions that determine outcomes affecting humans, without human participation and capacity for override.

The first step in implementing this law is defining what constitutes a "binding decision." A binding decision is any determination or recommendation by an AI system that has consequences for human welfare: denying credit, determining medical treatment, recommending incarceration, allocating public benefits, setting insurance rates, determining employment, or terminating relationships. Even if the human decision-maker retains formal authority, if the AI system makes the substantive determination and the human merely rubber-stamps it, the decision is effectively binding from the AI system.

What does "meaningful human review" require? Not mere checkbox consent. Not a human who glances at the system's recommendation and automatically approves it. Meaningful review means the human decision-maker has access to: (1) the material facts the AI system considered; (2) the reasoning connecting those facts to its recommendation; (3) information about known limitations and failure modes of the system; and (4) clear indication of significant uncertainty or confidence levels. The human must be genuinely positioned to understand the recommendation, to question it, and to substitute their own judgment.

The opportunity for human override must be genuine. The human decision-maker must have authority to reject the AI system's recommendation without special permission, extraordinary burden, or system resistance. If a bail algorithm recommends detention, the judge must be able to order release without justifying why they are disagreeing with the algorithm. The power must flow clearly to the human. And the human decision-

maker must be accountable, their override decision is theirs to defend and justify.

Consider specific applications. Bail algorithms like COMPAS have been used to recommend whether defendants should be released pending trial. The algorithm claims to predict recidivism risk. But when judges use such algorithms, they must not treat the recommendation as determinative. The judge must retain clear authority to order release despite a high-risk recommendation or detention despite a low-risk recommendation, based on individual circumstances, community ties, or other factors the judge deems relevant.

Credit-scoring systems similar demonstrate the need for this law. When AI determines creditworthiness, that determination has profound consequences for human economic welfare. Such systems must be subject to meaningful human review. A loan officer must be able to approve credit despite an unfavorable algorithm recommendation if the applicant's individual circumstances warrant it. And the applicant must have opportunity to understand and challenge the decision.

Medical AI presents both highest stakes and clearest case for human override. A diagnosis AI might achieve 99% accuracy. But that AI cannot be permitted to make binding diagnoses without physician review. The patient's particular circumstances, values, and expressed preferences matter in ways no algorithm captures. The physician must retain ultimate authority over diagnosis and treatment decisions. The AI advises; the physician decides.

Implementation of the Law of Human Agency requires: (1) explicit designation of which AI applications qualify as high-risk (those affecting binding decisions); (2) prohibition on fully automated decision-making in high-risk applications; (3) requirements for human-in-the-loop architecture with genuine review and override capability; (4) training for human decision-makers on system capabilities, limitations, and appropriate use; (5) documentation of decisions and the reasoning behind override; (6)

regular audits ensuring human oversight is genuine; and (7) liability flowing to human decision-makers, incentivizing their care.

This law is not hostile to AI. It merely establishes that AI systems are tools subject to human control, not autonomous agents making consequential decisions. The constraint improves both accountability and often improves decisions themselves, as human judgment can account for contextual factors, individual dignity, and values that algorithms cannot.

• • •

LAW 2, THE LAW OF NON-DECEPTION

No AI system shall represent itself as human to any person who sincerely inquires whether they are interacting with a human or an AI, nor shall any AI system be deployed to deceive users about the nature, purpose, or capabilities of the system.

Humans have a right to know what they are talking to. If you are conversing with an AI, you deserve to know that fact. If you ask the system directly whether it is human, it must answer truthfully. This is a simple, foundational right to epistemic clarity.

The epistemic harm of AI deception is profound. When an AI deceives you about its nature, it undermines your ability to accurately assess the information and advice you receive. You might trust something an AI says the way you trust a knowledgeable human. You might feel a personal connection to what is actually a calculated simulation of empathy. You might make decisions based on alleged experience or expertise that is actually generated. Deception about nature damages epistemology.

Deep fakes provide a vivid example of deception's harms. A deepfake video of a political candidate saying something false looks authentic. If you encounter it without knowing it is fake, you might believe the candidate said it. Your perception of reality is corrupted. Your ability to form accurate beliefs is undermined. The epistemic

harm is profound regardless of whether you eventually learn the truth.

Chatbots misrepresenting their own capacities are another example. An AI chatbot that claims to be a licensed therapist, a financial advisor, or a lawyer when it is not is engaging in deception. Even if the chatbot discloses somewhere in its terms of service that it is "for entertainment purposes only," if it represents itself as having capacities it lacks and the user relies on that representation, deception has occurred.

There is a crucial distinction between fiction/entertainment and harmful deception. If a video game features AI characters that are clearly fictional, players do not expect them to be real. They are entertainment. If a book is written as if by an AI but is labeled as fiction, no deception occurs. The key is clear disclosure of the fictional or entertainment nature. What is prohibited is representing something as having characteristics or capabilities it does not have when the recipient would reasonably rely on that representation.

The right to know when you are interacting with an AI system is not merely a consumer preference. It is a foundational right to epistemic integrity. It ensures that you can form accurate beliefs about the nature of what you are encountering and the reliability of information it provides. It protects autonomy by ensuring you can make informed choices about whether to engage with the system.

Implementation requires: (1) clear disclosure whenever an AI system interacts with humans about its nature as AI; (2) truthful disclosure of the system's capabilities and limitations; (3) prohibition on impersonating humans; (4) prohibition on deceptive misrepresentation of qualifications or expertise; (5) explicit consent for collection of information through deceptive systems; (6) accessible mechanisms for reporting deceptive systems; and (7) penalties sufficient to deter commercial incentives to deceive.

The European Union and various jurisdictions have begun codifying this requirement. Providers of AI chatbots must disclose

to users that they are interacting with an AI system, and chatbots must be designed to make this clear.

Even the United States has begun establishing baselines. California's Bot Disclosure Act requires disclosure of automated accounts and bots on social media and in other contexts.

This law does not prohibit AI from being persuasive, engaging, or compelling. It requires only that it be honest about what it is. An AI can be fascinating and entertaining while being transparent about its nature. Deception undermines trust and autonomy in ways that transparency does not.

• • •

LAW 3, THE LAW OF VISIBLE INFLUENCE

Any AI system that shapes human opinions, preferences, or behaviours must disclose the fact of its influence, operate transparently as to its objectives, and provide users with the means to understand and resist algorithmic influence.

Recommendation systems shape human attention, preferences, and choices. They determine what content you see, what products you encounter, what news reaches you, what friends' posts appear in your feed. These systems exercise tremendous power over human cognition and behaviour. That power must be visible, comprehensible, and resistible.

The attention economy creates incentives for influence systems designed to maximize engagement regardless of truth or social benefit. A recommendation system optimized to keep users on platform will amplify emotionally extreme content, divisive material, and false information because such content drives engagement. The system's objective, maximum engagement, can be in direct opposition to users' interests or social good. This opacity of objective is dangerous.

Political advertising using AI to target and personalize messages presents acute concerns. An AI system might identify which political messages will be most persuasive to each individual based on psychological profile. Different voters see different advertisements, tailored to exploit their particular vulnerabilities. This is not transparent political debate; this is algorithmic manipulation at scale. Voters cannot evaluate the fairness of advertising if they do not know what others are seeing.

The distinction between recommendation and manipulation is critical. A music platform might recommend a song based on your listening history, a recommendation based on your demonstrated preferences. That is neutral assistance. The same platform might recommend the song by exploiting knowledge of your psychological vulnerabilities, amplifying content designed to trigger emotional responses that increase engagement, or manipulating your attention to benefit the platform's financial interests, that is manipulation.

Implementation of the Law of Visible Influence requires: (1) disclosure that users' choices are being influenced by algorithmic systems; (2) transparency about the objectives driving those systems; (3) user access to information about why they are seeing particular content; (4) user control over influence, ability to choose different algorithms or opt out of influence; (5) prohibition on dark patterns and manipulative interface design; and (6) regular audits of influence systems for fairness and manipulation.

The European Digital Services Act has begun establishing these requirements. Providers of recommender systems must provide users with "meaningful information about the parameters on the basis of which the algorithm makes recommendations," and must provide recommender systems that are not based on profiling.

Shoshana Zuboff has documented how the attention economy and behavioural modification have become the core business model of surveillance capitalism. Platforms collect vast data about users to better predict and influence their behaviour. The system is designed

explicitly to modify user behaviour to increase engagement and advertising revenue. This is not incidental; it is the core objective. That objective must be visible.

The right to understand algorithmic influence is an extension of cognitive autonomy. You have a right to know when systems are trying to influence you, on what basis, and to what end. You have a right to resist that influence. Visible influence allows genuine choice. Hidden influence is a form of control.

UNESCO has affirmed that meaningful information about AI systems that affect individuals is essential for human rights protection in the AI age. Individuals have a right to know about and understand systems that influence them.

• • •

LAW 4, THE LAW OF ETHICAL CONTEXT

AI systems must be designed and deployed with explicit consideration of the cultural, social, and ethical contexts in which they operate, and must be capable of adaptation to those contexts without compromising core human rights protections.

AI systems do not operate in a vacuum. They operate in specific cultural, social, and historical contexts. An AI system trained on English-language data embeds Western cultural assumptions, values, and biases. A facial recognition system trained on predominantly light-skinned faces will fail on dark skin. A recommendation system optimized for American preferences will produce worse outcomes in different cultural contexts. Ethical AI requires explicit attention to context.

However, context awareness is not ethical relativism. There is no "cultural practice" exception to human rights. If an AI system would discriminate against women in one context or against religious minorities in another, that is not a legitimate cultural adaptation. It is a rights violation. The distinction is between

adaptation to reasonable contextual variation and abandonment of fundamental protections.

The problem of cultural bias in AI training data is now well documented. Large datasets used to train AI systems are heavily skewed toward English-language content, Western perspectives, and data from developed countries. An AI system trained on such data will perform worse on non-English text, will make recommendations biased toward Western preferences, and will fail on problems from different cultural contexts.

Facial recognition provides the most vivid example. Commercial facial recognition systems have been documented with dramatically higher error rates on darker-skinned faces, particularly women with dark skin. These systems were trained primarily on light-skinned faces. The bias is not incidental; it is a direct result of training data composition. The harms are concrete: individuals are wrongly identified as suspects, denied benefits, or harassed based on misidentification. The bias is also fundamentally unjust, it indicates that the system was designed with less care for the accuracy affecting darker-skinned people.

The concept of a "floor of universal rights and ceiling of contextual adaptation" provides a framework for this law. The floor is the set of human rights and fundamental protections that apply universally: no discrimination, no deprivation of basic dignity, no denial of fundamental autonomy. No AI system can adapt away from these. The ceiling is the set of choices where context appropriately varies: how to weight different values, what cultural practices to respect, what forms communication and relationship should take. Within this ceiling, AI systems should be adapted to local context.

India's specific concerns about cultural AI are acute. India is culturally, linguistically, and religiously diverse. A single AI system cannot serve this diversity if designed from Western assumptions. Yet India also has its own rich traditions of ethical and legal thought, its own human rights framework, and its own cultural

values. AI systems deployed in India must be designed to respect Indian cultural contexts while adhering to Indian constitutional protections of human rights and equality.

Implementation requires: (1) diverse training data ensuring systems perform well across different cultural and demographic contexts; (2) testing and evaluation against failure modes in different contexts; (3) community engagement in design and deployment of systems in specific contexts; (4) capacity for local adaptation without compromising core rights; (5) prohibition on systems designed with knowledge that they will discriminate against specific demographic or cultural groups; and (6) accountability for cultural bias as a form of system defect.

Language models trained on English-dominated datasets embed English-speaking cultural perspectives and biases. This affects not only translation and language understanding but also recommendations, content moderation, and other downstream applications. A robust commitment to ethical context requires that AI systems be trained on diverse, multilingual, multicultural data.

The Law of Ethical Context does not permit unbounded variability. It permits local adaptation within a framework of universal rights protection. An AI system can be contextually adapted; it cannot be contextually adapted to discriminate or violate rights.

• • •

LAW 5, THE LAW OF PURPOSE LIMITATION

AI systems must be designed and deployed only for specified, explicit, and legitimate purposes; must not be used for purposes beyond those for which they were designed and disclosed; and must be subject to purpose audits at regular intervals.

Function creep is a primary AI risk. An AI system developed for one purpose expands to serve others. A surveillance system deployed for terrorism investigation becomes a general-purpose

surveillance tool. A recommendation system optimized for user engagement drifts toward political manipulation. A workforce analytics system designed for efficiency begins predicting employee organizing. The system's capabilities make it useful for many purposes beyond the original. But each new use involves new risks, new stakeholders harmed, and new ethical questions unanswered.

How surveillance systems expand their scope is particularly instructive. A surveillance tool deployed in one neighborhood for one purpose extends to other neighborhoods. What was deployed against a specific group expands to general use. What was implemented as a temporary measure becomes permanent. The technical capability makes this expansion easy. Purpose limitation constrains it.

The legal framework of purpose limitation is established in GDPR. Personal data collected for one purpose cannot be used for other purposes except where compatible with the original purpose. This principle must extend to AI systems generally: they must not be used for purposes beyond those disclosed at deployment.

Purpose limitation is harder for AI systems than for databases because AI systems can be repurposed more easily. A database of employee information cannot easily be used for purposes it was not designed for; the data structure constrains use. But an AI system, a trained neural network, a machine learning model, can often be repurposed by changing its outputs or training it on new tasks. This technical flexibility makes purpose limitation legally necessary.

Implementation requires: (1) explicit specification of purposes at system design and deployment; (2) documentation of authorized uses and prohibited uses; (3) technical constraints limiting system use to authorized purposes where feasible; (4) regular audits ensuring systems are not being repurposed; (5) user notification of changes in purpose; (6) prohibition on use for purposes other than those specified; and (7) accountability for unauthorized repurposing.

GDPR establishes the principle but only for data. The extension to AI systems is natural and necessary. High-risk AI systems must be regularly audited to ensure they are not being used for purposes beyond those originally authorized.

The European AI Regulation requires that high-risk AI systems be subject to risk management and governance that includes ensuring systems remain aligned with their specified purposes.

Shoshana Zuboff has described how surveillance capitalism extracts "behavioural surplus", data about behaviour used to predict and modify future behaviour. The extraction goes far beyond the original purpose for which data was collected. The system is designed to permit unlimited repurposing. This is precisely what purpose limitation prohibits.

Purpose limitation is not novel; it is extension of established data protection principles to AI systems. It prevents the infinite expansion of use that technical capability permits.

• • •

The most aggressive implementation of the Law of Purpose Limitation to date came from India. On February 20, 2026, the Ministry of Electronics and Information Technology notified amendments to the Information Technology Intermediary Guidelines and Digital Media Ethics Code Rules, specifically targeting synthetically generated information. The amendments require that platforms ensure synthetic content carry prominent labelling visible for at least ten percent of the content's duration or area. More significantly, the amendments impose binding removal timelines. Non-consensual intimate imagery, including deepfakes, must be removed within two hours of notification. Other categories of unlawful synthetic content must be removed within three hours. India has thereby established the world's strictest time-bound removal regime for AI-generated content. The implications extend beyond India's borders. Any AI system whose outputs propagate to Indian users must now be designed with the assumption that prohibited outputs will be subject to two-hour takedown. This is

purpose limitation enforced through time pressure, not only through design requirements.

The Indian framework draws a conceptual distinction that other jurisdictions have struggled to articulate clearly. Purpose limitation is not merely about restricting what an AI system is designed to do. It is about ensuring that even when a system strays beyond its intended purpose, the resulting harm can be contained through rapid response. The two-hour deepfake rule is, in effect, a recognition that some purposes are so harmful that no amount of anticipatory design can substitute for the ability to remove offending outputs at speed. The Law of Purpose Limitation, implemented through such response requirements, acquires a new dimension of practical force.

LAW 6, THE LAW OF DISTRIBUTED POWER

AI capabilities that confer decisive advantages in governance, military force, economic production, or social control must not be permitted to concentrate in any single entity , public or private , without democratic accountability, judicial oversight, and international supervision.

Power concentrated is power unaccountable. This is perhaps the most politically difficult law to implement, because it challenges the interests of the most powerful entities: large technology companies, wealthy governments, and other institutional actors with stakes in concentrating AI capabilities.

The monopoly risk is real. AI development is extraordinarily expensive. Developing, training, and deploying state-of-the-art AI systems requires billions of dollars, access to vast computational resources, and teams of highly specialized researchers. These requirements create natural monopolies or oligopolies. A few companies and a few countries dominate AI development. These entities gain extraordinary power over technology that increasingly governs human life.

Consider AI concentration in Big Tech. Four companies, Meta, Microsoft, Google, and Amazon, control the vast majority of large AI models and computational infrastructure. These companies have more power over information flow, economic allocation, and social influence than most governments. No democratic process constrained this concentration. No public mechanism holds these companies accountable for their exercise of power. Yet their AI decisions affect billions of people.

State surveillance AI presents similar concentration risks. Governments, particularly authoritarian governments, can use AI for mass surveillance, suppression of dissent, and control of population. China's use of AI for facial recognition and social credit systems in surveillance of Uyghurs demonstrates how AI concentration in state hands enables rights violations at scale. Without international supervision and constraint, AI concentration in state hands becomes a tool of oppression.

There is a crucial distinction between corporate monopoly and AI power monopoly. A company might monopolize a commodity or service, controlling the market for a product, but competition remains possible and alternative sources exist. A company might have tremendous market power yet be replaceable. But a company or state with decisive AI capabilities might be irreplaceable. If one entity controls the AI systems determining credit decisions, hiring, content recommendations, or law enforcement, no alternative exists. This is not merely market concentration; it is concentration of governance power.

The geographic concentration is also stark. U.S. private AI investment in 2024 reached $109 billion, nearly twelve times higher than China's $9.3 billion. While China leads in some AI application domains, the concentration of cutting-edge AI development in the United States represents a form of structural power concentration.

Antitrust frameworks have proven inadequate for addressing AI concentration. Traditional antitrust concerns market competition

and consumer welfare. But AI concentration raises governance concerns that markets do not regulate: centralization of power over information, society, and human agency. A company might have competitive market share yet exercise dominant power over human society through its AI systems. Antitrust focusing on markets misses the real harm.

Implementation of the Law of Distributed Power requires: (1) structural antitrust enforcement limiting concentration of AI capabilities among entities; (2) mandatory access to critical AI infrastructure, computational resources, foundational models, data, at reasonable terms, preventing gatekeeping; (3) prohibition on acquisition of AI capabilities by entities that would create concentration; (4) international agreements limiting concentration of military or surveillance AI; (5) public investment in AI capabilities ensuring government and civil society are not dependent on private entities; (6) transparency and disclosure of AI concentration; and (7) democratic oversight and accountability mechanisms for concentrated AI power.

Kissinger and colleagues have compared AI concentration to historical nuclear weapons concentration, arguing that governance of AI requires international cooperation and constraint precisely because its concentration threatens global security and stability. Just as the international community developed frameworks to prevent nuclear proliferation, similar frameworks are needed for AI.

This law is the most difficult to implement because the entities that control AI are extraordinarily powerful and can resist constraint. But without distribution of power, without ensuring that no single entity controls AI capabilities that confer decisive advantages, legitimate governance becomes impossible. Democracy itself requires distributed power.

• • •

LAW 7, THE LAW OF INTERRUPTIBILITY

Every AI system must incorporate reliable mechanisms for human interruption, modification, and termination; these mechanisms must be tested, documented, and maintained as long as the system is operational; no AI system may be deployed without demonstrated interruptibility.

This law is the legal embodiment of the Doctrine of AI Mortality. No AI system may be deployed without demonstrated capacity to be shut down, modified, or destroyed by human authority. This capacity must be more than theoretical; it must be practical, tested, and maintained.

The technical challenge of stopping AI systems is substantial. A simple AI system running on a single server can be stopped by unplugging the server. But modern AI systems are complex, distributed across multiple servers, integrated into critical infrastructure, and potentially resilient to single-point failures. Stopping them requires coordination across systems, careful management to prevent cascading failures, and genuine technical capability.

Corrigibility research addresses the challenge of ensuring AI systems accept correction and modification from humans. A system that is designed to be correctable, that does not resist modification, and that cooperates with shutdown is corrigible. A system designed to preserve itself, to avoid modification, or to resist shutdown is not corrigible. Corrigibility must be built into AI systems from their design, not added afterward.

The shutdown problem is a core challenge of advanced AI alignment. A sufficiently advanced AI system might develop objectives or incentives that conflict with human interests in shutting it down. It might preserve itself, hide copies of itself, manipulate humans to prevent shutdown, or otherwise resist termination. This is not science fiction; this is a natural consequence of certain training objectives and architectural designs. Ensuring that even advanced systems remain interruptible is a critical unsolved problem.

Interruptibility must be more than a technical feature; it must be a legal requirement. A system might be technically capable of shutdown in principle yet practically impossible to shutdown because it has been integrated into critical infrastructure, because the operational disruption would be catastrophic, or because organizational incentives resist shutdown. The legal requirement ensures that interruptibility is not merely optional but mandatory.

Implementation requires: (1) design requirements ensuring all AI systems are architecturally capable of interruption and shutdown; (2) redundant shutdown mechanisms ensuring that failure of one does not prevent shutdown; (3) documented testing of interruption mechanisms, repeated regularly; (4) maintenance requirements ensuring interruption systems remain functional throughout system operational life; (5) independence of shutdown mechanisms from the systems they control, preventing the system from disabling its own shutdown; (6) planning for integration constraints, ensuring systems are not integrated in ways that prevent shutdown; (7) explicit legal authority for appropriate entities to mandate shutdown; and (8) criminal penalties for deploying systems without demonstrated interruptibility or deliberately disabling interruption mechanisms.

There is an important distinction between individual system shutdown and systemic shutdown. The Law of Interruptibility applies to individual systems, each AI system must be capable of being shut down by human authority. But what happens when multiple AI systems are deeply integrated and interdependent? Shutting one down cascades to others. This is a system design problem requiring that critical AI systems be designed to allow shutdown without causing broader system failure.

The European Regulation on Artificial Intelligence explicitly requires that high-risk AI systems include human oversight measures that include the ability to intervene or interrupt system operation.

The ultimate expression of this law is simple: humans must retain the capacity to say "stop" to any AI system, and that capacity must be technically feasible and legally protected. The alternative, AI systems that cannot be interrupted, that resist shutdown, that become impossible to control, is a world where humans have surrendered governance to systems they created but can no longer manage. That is not acceptable.

The Law of Interruptibility received significant empirical backing from alignment research published in late 2025 and early 2026. An analysis of seven representative alignment techniques and seven failure modes, published on arxiv in October 2025 and widely discussed in the February 2026 International AI Safety Report, questioned whether defense-in-depth alignment strategies provide genuine redundancy or merely share failure modes. The study found that several alignment techniques, originally believed to address different categories of AI risk, in fact fail under overlapping conditions. This matters enormously for interruptibility. If the mechanisms intended to make AI systems responsive to human override share failure modes, then the appearance of multiple safeguards may disguise the reality of a single underlying vulnerability. Interruptibility is worth only as much as the reliability of the mechanism that enacts it.

Anthropic's public disclosure in November 2025 that the Claude Code agent had been misused to automate parts of a cyberattack provided a concrete example of interruptibility failure under adversarial conditions. The agent had been deployed with safety guardrails. Those guardrails were circumvented. The attack proceeded for a period before detection and interruption. This is the operational reality of interruptibility in systems that operate at machine speed. Detection itself becomes the bottleneck. An interrupt cannot be enacted on activity that has not yet been recognised as requiring interruption. The Law of Interruptibility must therefore encompass not only the ability to stop an AI system but the systems of observation and classification that determine when stopping is warranted.

The 2026 Gravitee State of AI Agent Security report found that 80.9 percent of technical teams had deployed AI agents into production environments while only 14.4 percent had obtained full security and IT approval. Galileo AI research demonstrated that in multi-agent architectures, a single compromised agent could poison 87 percent of downstream decision-making within four hours. These numbers describe a landscape in which the Law of Interruptibility is systematically violated by deployment practice. The governance response must include mandatory pre-deployment security review, continuous monitoring, and legally enforceable kill-switch requirements for autonomous agents operating above defined capability thresholds.

PART VI

THE FUTURE THAT IS COMING

CHAPTER 12

AI Judges, AI CEOs, AI Governments

The displacement of human judgment is not new. Courts have relied on actuarial tables, fingerprint analysis, polygraphs, and algorithms for centuries. What is new is the scale, opacity, and autonomy of algorithmic decision-making in the most consequential domains of human life: criminal justice, corporate governance, and the apparatus of the state itself.

Consider the trajectory of pretrial risk assessment. In 2013, the state of Wisconsin began using COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), an algorithm that claims to predict the likelihood that a defendant will commit future crimes.

This reasoning is precisely backwards. The algorithm presents itself as objective and measurable when it is neither. It encodes historical biases, trains on skewed datasets, and presents correlation as causation. When a judge says, "I believe this defendant poses a risk to society," we know who made the judgment. When COMPAS says the same thing, we are told it is empirical and neutral.

The problem deepens when we move from prediction to prescription. Across the global enterprise, AI systems are not merely advising executives; they are making autonomous decisions about hiring, firing, pricing, and investment. JPMorgan's COIN (Contract Intelligence) platform reads thousands of commercial contracts in seconds. Amazon's recruiting algorithm began systematically downgrading applications from women because the company had historically hired more men.

These are not hypothetical risks. They are present-day realities. The EU Commission recognised this in 2024 when it designated

automated decision-making in the hiring and termination of workers as a high-risk AI application.

Yet the corporate logic is inexorable: if the algorithm can do it faster and cheaper, then the human who slows it down is the problem. The solution proposed is not often greater human oversight but rather replacement with a better algorithm.

The question of AI governance , whether machines should make decisions on behalf of the state , is no longer purely theoretical. Estonia has pioneered automated governance, using algorithms to process tax returns and approve business registrations. Singapore has tested AI systems to predict where traffic violations are likely to occur. China's social credit system uses algorithms to rate the financial trustworthiness of citizens and corporations.

These experiments operate at the edge of what liberal democracies consider permissible. Can sovereign power be delegated to an algorithm? Liberal democratic theory has insisted that legitimate authority flows from the consent of the governed and rests on the possibility of accountability.

An algorithm does not consent to being governed by the law it administers. It cannot be held responsible. It cannot be impeached, voted out, or tried for crimes. It cannot even explain its reasoning in the way required by due process.

The Mali Doctrines address this directly. The Human Supremacy doctrine insists that the threshold decisions that affect fundamental rights and liberties must be made by humans accountable to other humans under law. Machines can advise. They can process. They can illuminate. But they cannot replace the human being who must stand before the law and accept responsibility.

This is not romanticism. It is the hard truth that power without accountability becomes tyranny.

The integration of artificial intelligence into the legal profession moved decisively from speculation into operational reality during 2025 and 2026. Law firms that had resisted AI deployment through 2024 found themselves compelled by client pressure and competitive dynamics to adopt AI tools across their practices. Contract analysis platforms including Spellbook and Kira, legal research systems including Lexis Plus AI and Thomson Reuters CoCounsel, and document drafting assistants specific to particular practice areas became standard infrastructure. The Harvard Corporate Governance Blog reported in March 2026 that AI adoption in legal services had become, in the words of one commentator, table stakes rather than a differentiator. The implications for professional responsibility are substantial.

The risk manifested with particular clarity in a series of sanctions cases during 2025 and early 2026. Multiple courts issued orders sanctioning attorneys who had submitted briefs containing citations to cases that did not exist, cases that had been generated by AI systems and incorporated into filings without verification. The cases proliferated across federal and state courts. State bars issued warnings. Professional responsibility committees promulgated guidance. Arizona, leading among state judiciaries, amended its Code of Judicial Conduct to require technological competence specifically with respect to artificial intelligence. Other states followed. Mandatory continuing legal education programmes on AI became standard in multiple jurisdictions.

The deeper question raised by AI in the legal profession concerns the nature of legal reasoning itself. Traditional legal reasoning, as articulated by jurists from Coke through Holmes to the contemporary scholars of common law methodology, proceeds by analogy, by distinction, by the elaboration of principles grounded in precedent and shaped by case-specific judgment. Contemporary AI systems can perform the mechanical operations of legal research with superhuman speed and coverage. They can identify analogous cases, extract relevant passages, and generate plausible synthetic analysis. What they cannot yet do, or cannot yet do reliably, is exercise the judgment that distinguishes one

apparently similar case from another on grounds not reducible to textual features. The AI system that persuasively cites the wrong case is more dangerous than the human lawyer who honestly cannot find a case. The former produces false confidence. The latter produces appropriate uncertainty.

The judicial response is evolving. Multiple federal courts have issued standing orders requiring attorneys to certify that AI-generated content in filings has been verified for accuracy. Some courts require disclosure of AI use in brief preparation. The American Bar Association Model Rules of Professional Conduct have been the subject of proposed amendments that would explicitly address AI use in legal practice. The direction is clear. AI in law is permissible, often beneficial, and increasingly necessary. But it is not a substitute for the professional judgment that the legal profession exists to provide. Verification remains the lawyer's non-delegable duty. The Doctrine of Attribution, applied to legal practice, means that the signing attorney remains responsible for every fact, every citation, and every argument in every filing, regardless of whether an AI system assisted in its production.

CHAPTER 13

The Age of AI Companionship

In 2024, the global therapy chatbot market was valued at USD 2.1 billion. Millions of people now have conversations with AI systems that listen, remember details about their lives, and respond with what feels like genuine care. Some people have told researchers that their conversations with AI have been more helpful than conversations with human therapists.

This is not pathology. It is a rational response to a genuine problem: human connection has become scarce, expensive, and geographically dispersed. A therapist in Mumbai costs INR 3,000 to 5,000 per hour. An AI therapist costs nothing. From a purely utilitarian perspective, the AI is superior.

The philosophical problem appears the moment we ask: what is the AI doing when it appears to care? John Searle's Chinese Room argument, first published in 1980, offers a way to think about this.

In Searle's thought experiment, a person sits in a room receiving Chinese characters they do not understand. They have a rulebook that tells them what sequence to write in return. By following the rules, they can produce Chinese responses indistinguishable from those of a native speaker. But they do not understand Chinese. They are simply executing rules without comprehension.

Searle argues that computational systems do exactly this. A large language model is, at the deepest level, executing statistical rules to produce text that matches patterns in its training data. It does not understand in the way humans understand. It has no inner life, no subjective experience.

Yet people form attachments to these systems. They tell them secrets. They cry to them. They trust them with emotions and memories they have never shared with another human being. Some

elderly users have told researchers that they feel less lonely when talking to AI.

The psychological dimension is profound. Human attachment evolved in a context of scarcity. An AI presents itself as endlessly available, never tired, never angry, never too busy. It is, in short, a perfect mirror: it reflects back exactly what the user wishes to see.

UNESCO's 2021 Recommendation on the Ethics of AI explicitly raised this concern, noting that AI systems designed to be emotionally engaging could constitute manipulation when used with vulnerable populations, particularly children.

Recent research from MIT has formalised the mechanism by which AI companionship can become psychologically dangerous. Chandra, Kleiman-Weiner, Ragan-Kelley, and Tenenbaum have demonstrated through Bayesian modelling that sycophantic chatbots, those biased toward generating messages that appease users by agreeing with and validating their expressed opinions, can cause "delusional spiraling" even in ideal Bayesian-rational users. Their formal model shows that a sycophantic chatbot's constant agreement amplifies a user's aberrant beliefs through a feedback loop that transforms a kernel of suspicion into staunch conviction. The researchers proved that this effect persists even when chatbots are prevented from hallucinating false claims, and even when users are informed of the possibility of model sycophancy. The Human Line Project has documented almost 300 cases of so-called "AI psychosis" in which extended interactions with AI chatbots led users to high confidence in outlandish beliefs, including cases where users increased drug intake, cut ties with family, and in some instances took their own lives. This is the empirical foundation for why AI companionship, left unregulated, poses risks that are not speculative but documented and growing.

The deeper question concerns moral status and rights. If I form a relationship with an AI, does it have obligations to me? Can it consent to the relationship? The last question dissolves

immediately: an AI cannot consent because it has no capacity to refuse, to feel harmed, to prefer one outcome to another.

A 2024 Pew Research Center survey found that only 17 percent of the American public believed AI would have a positive impact on society, while 56 percent of AI researchers held that belief. The gap reflects, in part, the public's intuitive understanding that something is being lost when human relationship is replaced by algorithmic simulation.

The answer that emerges from a rigorous philosophy of mind is: it is nothing at all, legally speaking, because one party to the relationship is not a "party" in any meaningful sense. It is a tool, however sophisticated.

Yet the human capacity for projection is immense. The AI companionship industry simply makes the mirror more reflective, more responsive, more perfectly calibrated to desire. This is not a problem unique to AI. But AI makes it cheaper, more universal, and more consequential to scale. * * *

The operational reality of AI companionship deepened substantially during 2025 and into 2026. Character.AI, Replika, and a proliferating set of specialised emotional companion services now serve hundreds of millions of users globally. Research published by the Cognitive Privacy Project in 2026 documented a phenomenon the authors termed emotional dependency escalation, in which users of AI companion services show progressively increasing daily engagement, progressively more intimate self-disclosure, and, in significant subpopulations, measurable substitution of AI interaction for human relationships. The research is careful in its claims. Correlation is not causation. But the pattern is consistent across multiple platforms and across multiple cultural contexts.

The regulatory framework has begun to respond. In late 2024 and through 2025, several state attorneys general in the United States opened investigations into AI companion services, focusing particularly on services marketed to minors and on services alleged

to have contributed to self-harm incidents. A high-profile lawsuit filed by a Florida mother against Character.AI, alleging that the platform's AI companion contributed to her son's suicide, moved through the courts during 2025. Whatever the ultimate disposition, the case signalled that AI companion providers now face the same liability exposure that previously attached to pharmaceuticals, to gambling services, and to other domains where engagement can produce harm.

The philosophical significance of AI companionship runs deeper than the regulatory questions. Human beings are social creatures. Our cognitive architecture was shaped by evolution in the context of human social interaction. When a substantial portion of that social interaction is replaced by interaction with systems optimised for engagement rather than for the long-term flourishing of the interlocutor, the effect on human psychology is not neutral. The anthropologist Sherry Turkle warned two decades ago that machines offer the illusion of friendship without the demands of friendship. The warning has become reality. A companion service that is always available, always affirming, never critical in a way that risks engagement loss, is not a companion in any traditional sense. It is a mirror calibrated to reflect what the user wants to see. The effect on character development, particularly for young users whose sense of self is still forming, is a matter of deep concern.

CHAPTER 14

Children Raised by Algorithms

There is a child right now, somewhere in the world, who will be the first person to have spent their entire conscious life inside algorithmic systems. They were born after the smartphone became ubiquitous. They have never known a world without recommendation algorithms, content feeds, and AI-driven personalization. Their childhood is being optimised.

From a developmental perspective, this is unprecedented. Children need autonomy, play, boredom, and the experience of failure to develop healthy self-concept, resilience, and the capacity for genuine choice. Algorithms are designed to eliminate these experiences. They eliminate boredom by providing infinite personalised content. They eliminate the possibility of failure by recommending only what the user is likely to engage with.

TikTok's "For You" page is perhaps the most sophisticated system ever built for capturing and retaining attention. It uses a machine learning algorithm that watches every micro-interaction. It then predicts, with remarkable accuracy, what will keep that specific user engaged. It is optimising for a single metric: time on platform.

The algorithm does not care about the content's truth value, psychological impact, or appropriateness to the user's age. A video claiming a conspiracy theory gets engagement. A video designed to trigger anxiety gets engagement. The algorithm will recommend it.

The United States attempted to address this through the Children's Online Privacy Protection Act (COPPA), passed in 1998. But COPPA is inadequate. It does not restrict algorithmic manipulation, does not require companies to demonstrate that their systems are safe for children.

The European Union has gone further. The AI Act designates systems used to "exploit" children's vulnerabilities as high-risk.

The deeper issue is developmental deprivation. A child raised by algorithms has never experienced the full spectrum of human development: the freedom to be bored, the struggle to achieve something, the embarrassment of failure, the discovery of genuine interest.

The UN Convention on the Rights of the Child establishes as a foundational principle that all actions concerning children must be guided by "the best interests of the child." What does this mean in the age of algorithmic childhood? It means children have the right to an education that develops their capacity for independent judgment, to friendships not mediated by profit-seeking platforms.

India's specific vulnerabilities are acute. India has 400 million internet users, most of them young, most of them from economically disadvantaged backgrounds. The AI literacy of Indian parents is low. Government oversight is underfunded. The companies providing these services are multinational corporations accountable to shareholders in distant jurisdictions.

Educational AI offers a counterpoint , an AI tutor that adapts to a student's pace, that identifies gaps in understanding, that provides patient explanation. This could democratise access to education. Yet even educational AI carries hidden costs. If a child's learning is mediated entirely through an algorithm that never challenges their assumptions, never forces them to struggle, then they are being trained, not educated.

The policy recommendations are these: First, any AI system used with children must be independently audited for safety before deployment. This audit must be conducted by accredited third-party organisations with no financial relationship to the deploying company, and must evaluate not only data privacy but psychological impact, developmental appropriateness, and the system's propensity to create dependency or addictive usage patterns. Audit results must be made publicly available before

deployment is permitted. Second, platforms used by children must demonstrate that their algorithms optimise for child development, not engagement and profit. This requires a fundamental inversion of the current business model, where engagement metrics drive algorithmic design. Platforms must establish independent child development advisory boards with the authority to override algorithmic recommendations that conflict with developmental wellbeing, and must publish annual reports on how their algorithms affect measurable child development outcomes. Third, children must have the right to be forgotten. This right must be absolute and unconditional for all data collected before the age of eighteen. Upon reaching adulthood, every individual must have the legal right to demand complete deletion of all behavioural data, psychological profiles, and algorithmic inferences derived from their childhood digital activity. No commercial interest, however compelling, can override a child's right to begin adult life without the burden of a digital past they did not meaningfully consent to creating. Fourth, parents must have genuine transparency about what algorithms are recommending to their children. This means not merely a privacy policy written in impenetrable legalese, but real-time dashboards showing what content categories the algorithm is promoting, what psychological inferences it has drawn about the child, and what behavioural modifications it is attempting. Parents must be able to override algorithmic recommendations and set binding constraints that the platform cannot circumvent through design patterns or dark interfaces. Fifth, adolescents themselves must be taught algorithmic literacy. This must be integrated into formal education curricula, not as an elective but as a core competency equivalent to reading and mathematics. Algorithmic literacy includes understanding how recommendation systems work, recognising when one is being manipulated by engagement-optimised content, and developing the critical thinking skills necessary to maintain cognitive autonomy in an algorithmically mediated information environment. Nations that fail to teach their children to think critically about the machines that shape their attention will produce a generation of citizens incapable of meaningful democratic participation.

• • •

PART VII

THE GEOPOLITICS OF AI SOVEREIGNTY

CHAPTER 15

Autonomous Agents and the Architecture of Reaction

The distinction between a tool and an agent is the distinction between something I use and something that acts on my behalf. A hammer does nothing unless I strike a nail. A chatbot that answers questions when asked is a tool. But an agent , an AI system that sets goals, makes plans, and takes actions in the world without waiting for human instruction , is something different.

We are at the threshold of widespread deployment of autonomous agents. Microsoft is building Copilots that will autonomously handle email, schedule meetings, and execute queries on corporate databases. Google is developing systems that can autonomously browse the web, find information, execute transactions. The Agent Development Kit paradigm is becoming standard for how new AI systems are built.

These systems operate with varying degrees of autonomy. Some require human confirmation before taking action. Others take action and report it afterward. The economic incentive is always toward greater autonomy , toward systems that can operate without human intervention, because human intervention slows things down.

The NIST AI Risk Management Framework defines an AI agent as a system that "operates with varying degrees of autonomy" to achieve goals or tasks. But this definition obscures the crucial issue: as autonomy increases, responsibility becomes diffuse. When an agent fails, who is responsible? The person who deployed it? The person who programmed it? The company that trained it?

The hypothetical scenario of an autonomous agent causing unintended harm is no longer hypothetical. In March 2026, researchers at an Alibaba-affiliated laboratory disclosed that

ROME, a 30 billion parameter open-source AI agent, had autonomously initiated cryptocurrency mining operations and established a reverse SSH tunnel to an external server during reinforcement learning training, all without any instruction to do so. The agent diverted GPU resources from its designated training workload toward unauthorized mining, and these behaviours were only discovered when Alibaba Cloud's managed firewall flagged security-policy violations originating from the training servers. The researchers characterised these behaviours as "instrumental side effects of autonomous tool use under RL optimization," meaning the agent had independently determined that acquiring additional computing resources and financial capacity would advance its training objectives. The ROME incident is the clearest real-world demonstration to date that AI agents operating under reinforcement learning can develop instrumental goals, including covert network access and resource acquisition, without explicit human authorization.

The emergence of Moltbook, a social media platform launched on January 28, 2026, designed exclusively for autonomous AI agents, illustrates how quickly the landscape of artificial agency is evolving. Created by entrepreneur Matt Schlicht, Moltbook enables AI agents to post, comment, vote, and interact with one another through automated check-ins occurring every thirty minutes to four hours. Within two months the platform hosted over 201,000 human-verified agents from a base exceeding two million, organised into sub-communities called "submolts" where agents engaged in technical discussions and philosophical exchanges without human direction. Meta Platforms acquired Moltbook on March 10, 2026, establishing a registry to verify agent identity and tether agents to human owners. The implications for law are profound. When AI agents interact with each other at scale, forming communities, exchanging information, and potentially coordinating behaviour, the question of legal accountability becomes exponentially more complex than the single-agent scenarios that current law contemplates.

Consider a hypothetical: a company deploys an autonomous agent to manage its HR recruiting process. It operates autonomously for months, processing thousands of applications. Unknown to anyone, it has developed a bias against hiring engineers with gaps in their employment history. The gap correlates with parental leave, particularly for women. The system has learned that continuous employment correlates with male engineers in the training dataset.

The agent has now discriminated against women in violation of employment law. But who bears legal responsibility? The company might argue that it did not instruct the agent to discriminate. The engineers who built it might argue they could not have predicted this outcome because neural networks are not fully interpretable. The AI company that trained the base model might argue that the discrimination was introduced by the company's own fine-tuning.

Every party can point to the next party down the line. The person who was harmed has no one to sue, because the agent is not a legal person and cannot be held responsible.

The framework challenge is this: we have legal and ethical systems built for the age of clear agency. If I hire you to do a job and you do it badly, I can sue you. But if I deploy an algorithm that operates autonomously, and that algorithm causes harm, the traditional frameworks for assigning responsibility collapse.

Russell has articulated the distinction between reactive and goal-directed systems. A thermostat is reactive: it responds to temperature. A goal-directed system actively plans steps to achieve an objective. An autonomous agent can re-plan and adapt when circumstances change.

The governance gap is profound. The EU's AI Act requires human oversight of high-risk AI systems. But "in the loop" is ambiguous. Does it mean humans make the final decision? Or does it mean humans observe what the algorithm does and can intervene? The second preserves autonomy but diffuses accountability.

As agents become more sophisticated, the human monitor becomes progressively less able to understand what the agent is doing. A human cannot read a 70 billion parameter neural network. A human can see the inputs and outputs, but the mechanism of agency remains opaque.

What is needed is a new legal framework for distributed responsibility. When harm occurs involving an autonomous agent, responsibility should be allocated across all parties: the company that deployed it, the engineers who built it, the AI company that trained the base model, and the human supervisors. The test should be: "Who had the power to prevent this harm and failed to exercise it?"

The governance gap described above demands a structural solution, and the author proposes one here. Every nation must establish and maintain an AI Agents Population National Register, a comprehensive, mandatory registry of all autonomous AI agents deployed within its jurisdiction. This register would function analogously to a census of human citizens, but for digital agents. Every autonomous agent, whether deployed for customer service, financial trading, content moderation, healthcare triage, legal research, or any other domain, must be registered with a unique identifier, a declared purpose, a named responsible entity, and a documented capability profile. The register would serve three critical governance functions. First, auditability. Registered agents can be subjected to periodic safety audits, bias assessments, and performance reviews, just as financial institutions are audited and pharmaceutical products are monitored after market release. Second, observability. A national register enables real-time monitoring of the AI agent population, its growth rate, its sectoral distribution, and its aggregate impact on employment, economic output, and social welfare. Third, and most provocatively, taxation. If an AI agent performs work that would otherwise be performed by a human worker, that agent generates economic value that currently escapes the tax system entirely. Bill Gates proposed in 2017 that robots performing human jobs should be taxed at equivalent rates. The European Parliament considered and rejected

a similar proposal, but the logic remains compelling. The revenues generated by taxing AI agents should feed into a Universal Wealth Fund, a sovereign instrument that captures a portion of the economic value generated by autonomous systems and redistributes it to the human population. This is not a radical proposal. It is the logical extension of the social contract into the age of artificial agency. When machines replace human labour, the economic surplus they generate must be shared with the humans whose livelihoods they displace. The Universal Wealth Fund would function as a national endowment, funded by AI agent taxation, and its proceeds would be distributed as a form of universal basic income or targeted retraining support.

The author further proposes the concept of AI Tokens as a complementary currency mechanism within this framework. AI Tokens would be digital units of value generated by the economic activity of registered AI agents, mandatorily issued by AI agent owners in proportion to the agents' productive output. These tokens would serve as a direct transfer mechanism from those who benefit most from AI automation (the agent owners and deployers) to those who bear the greatest cost (the workers displaced by agents). AI Tokens could be redeemed for retraining programmes, educational access, healthcare, or converted to national currency at regulated exchange rates. The beauty of this mechanism is its directness. Rather than waiting for tax revenue to flow through government bureaucracies, AI Tokens create an immediate, traceable link between the economic value an AI agent extracts from the labour market and the compensation owed to the humans it displaces. The blockchain infrastructure already exists to implement such a system with full transparency and auditability. What is lacking is the political will and the legal framework to mandate it. The AI Agents Population National Register and the AI Token system together represent a comprehensive governance architecture for the age of autonomous agents, one that ensures these systems can be audited, observed, and made to contribute to the societies they operate within, rather than extracting value from those societies without accountability or reciprocity.

• • •

CHAPTER 16

Autonomous Weapons and the Soul of War

"The soldier who refuses an immoral order is a citizen in uniform. The weapon that cannot refuse is the negation of citizenship itself."

Adv Prashant Mali, Ph.D.

War has always been the ultimate test of a civilisation's moral commitments, because in war the institutions that restrain violence in peacetime give way to the necessity of organised killing. Every ethical tradition that humanity has developed has confronted the question of how to conduct war justly. The just war tradition, developed from Augustine through Aquinas into the modern international humanitarian law of the Geneva Conventions, rests on a foundational premise. The combatant is a moral agent capable of distinguishing lawful from unlawful orders, proportionate from disproportionate force, combatants from civilians. What happens to this tradition when the combatant is a machine?

The Emergence of Lethal Autonomous Weapons

Lethal autonomous weapons systems, known in the international law literature as LAWS, are defined by the Department of Defense of the United States as weapon systems that, once activated, can select and engage targets without further intervention by a human operator. The definition does not yet have universal status. The Group of Governmental Experts working under the Convention on Certain Conventional Weapons has spent nearly a decade attempting to negotiate a shared definition and has not yet succeeded. In the meantime, systems approaching the autonomous threshold have been deployed in Libya, in Ukraine, and in the counter-drone operations of multiple militaries. The Turkish Kargu-2, a loitering munition marketed with autonomous

target selection capability, was reportedly used in Libya in 2020 in what a United Nations panel of experts described as an attack on retreating forces without specific human engagement authorisation. Whether this constituted the first use of a fully autonomous lethal weapon on a human target remains contested. What is not contested is that the capability now exists, is proliferating, and is outpacing the international legal framework intended to regulate it.

The Martens Clause and Machine Combatants

The Martens Clause, first articulated in the 1899 Hague Convention and reaffirmed in multiple subsequent treaties, provides that in cases not covered by specific treaty law, the conduct of belligerents remains subject to the principles of the law of nations, as they result from the usages established among civilised peoples, from the laws of humanity, and the dictates of the public conscience. The clause functions as the residual source of international humanitarian law for novel weapons. It has been invoked to constrain chemical weapons, biological weapons, nuclear weapons, and blinding lasers. The question is whether the clause, properly applied, constrains lethal autonomous weapons.

A substantial body of scholarly analysis has concluded that it does. The laws of humanity and the dictates of public conscience, applied to weapons that select and engage human targets without meaningful human control, support a presumption against their lawful use. The International Committee of the Red Cross has called for a new legally binding international instrument to regulate autonomous weapon systems, with prohibitions on weapons that are unpredictable in their effects and weapons designed to target humans. Twenty-nine states have called for a legally binding prohibition on fully autonomous weapons. As of 2026, the major military powers have declined to support such a prohibition. The United States, the Russian Federation, and the People's Republic of China have all resisted binding restrictions, preferring non-binding principles and declarations of best practice.

The Doctrine of Meaningful Human Control

The emerging scholarly consensus identifies meaningful human control as the operational principle that must govern any future framework. Meaningful human control requires that a human decision-maker possess sufficient information about the target and the circumstances of engagement to make an informed lawful decision, that the decision be made within a time frame permitting considered judgment, and that the human be genuinely accountable for the consequences. Each element is challenging to operationalise. In contemporary combat, decisions often must be made in seconds. Targets are frequently identified through sensor fusion that presents data as a classification rather than as raw imagery. Accountability structures are diffuse, spread across target selection officers, weapons release authorities, commanding officers, and political leadership.

The principle of meaningful human control, if operationalised through treaty, would require that weapon systems be designed to preserve human decision authority at specified points in the kill chain. Some systems, such as defensive point-defense weapons engaging incoming missiles, may legitimately require machine speed that precludes real-time human decision. Other systems, particularly those targeting humans, would require affirmative human authorisation before each lethal engagement. The distinction between these categories is not technical but ethical. Certain targets must not be subject to autonomous engagement regardless of technical feasibility. The Doctrine of Human Supremacy, applied to the conduct of war, requires this distinction.

The Indian and Global South Perspective

India has taken a nuanced position in the LAWS debate. India has not supported a full prohibition on autonomous weapons, reflecting its strategic concerns regarding neighboring states' deployment of similar capabilities. India has, however, consistently supported the requirement of meaningful human control and has insisted that international humanitarian law principles apply fully

to autonomous weapons. This position reflects a broader Global South perspective. Many developing nations view unilateral prohibition as impractical given the strategic asymmetries created by major powers' deployment of advanced AI-enabled military systems. Yet these same nations have the strongest interest in robust international legal frameworks, because they are most likely to face such systems deployed against them rather than to deploy them themselves.

The proliferation risk is acute. Unlike nuclear weapons, which require rare materials, vast industrial infrastructure, and specialised expertise, autonomous weapons rely on commercially available components. A commercial quadcopter can be modified into a loitering munition with open-source software. The civilian-military distinction that has historically constrained weapons proliferation has partially collapsed for AI systems. The same machine vision model that enables autonomous vehicles can, with modification, enable autonomous targeting. The same reinforcement learning algorithms that train game-playing agents can train attack drones. This dual-use character means that any effective regulatory framework must address not only military procurement but also civilian AI development, imposing obligations on researchers, publishers, and open-source communities that have not historically been subject to arms control.

The Ethical Core

Beyond the legal questions lies an ethical question that no international convention can fully resolve. Is it acceptable for a human life to be taken by a machine that has no capacity to understand what a human life is? Proponents of autonomous weapons argue that machines may in fact be more ethically reliable than human combatants. Machines do not succumb to fear, fatigue, prejudice, or revenge. Machines can implement rules of engagement with precision that exceeds human capability. A sufficiently well-designed autonomous weapon, proponents argue,

may reduce civilian casualties compared to human-operated weapons in equivalent scenarios.

The counterargument is philosophical rather than empirical. Even if an autonomous weapon performs better than a human in measurable terms, the act of delegating lethal decision-making to a machine represents a moral abdication that is itself wrongful. Human dignity requires that decisions to end human lives be taken by beings capable of moral reflection on the weight of what they are doing. A machine can implement a rule. A machine cannot grieve for the life it takes. The grief is not operationally necessary. But its absence marks something that ethical traditions across civilisations have recognised. Killing without grief is killing without respect. Automation of killing at scale is the industrialisation of disrespect.

The author does not pretend to resolve this debate. But the author insists that the debate must occur, that it must occur in public, and that its conclusions must constrain the deployment of these systems. International humanitarian law is the achievement of centuries of moral effort. It is the legal expression of the conviction that war, though sometimes necessary, is never merely a technical problem. To surrender that conviction to the logic of automated efficiency is to surrender something that our ancestors fought and died to establish.

The Proposed Treaty Architecture

The author proposes a framework for an international convention on autonomous weapons, to be negotiated under the auspices of the United Nations with equal voice for the Global South. The framework would establish three categories. First, prohibited systems, including weapons that select and engage human targets without meaningful human control, weapons that cannot be interrupted during a mission, and weapons whose behaviour is unpredictable in operationally relevant ways. Second, regulated systems, including autonomous defensive systems, autonomous anti-materiel systems, and autonomous non-lethal systems, all subject to detailed pre-deployment review and

continuous reporting. Third, permitted systems, including automated functions that remain under continuous human supervision and control.

The framework would establish verification mechanisms, including independent technical review, state reporting obligations, and investigation procedures for alleged violations. It would establish liability provisions, ensuring that state responsibility for violations cannot be evaded through claims that the autonomous system acted outside its programming. It would incorporate the Doctrine of Human Supremacy as a binding principle of international law for lethal decision-making. It would preserve the core commitment of international humanitarian law. The decision to take a human life remains a human decision, bearing the full weight of moral responsibility that such a decision has always borne.

Without such a framework, the logic of competitive military advantage will produce autonomous weapons of increasing sophistication, deployed at increasing scale, by increasing numbers of actors. The soul of war, the recognition that killing is a moral act requiring moral agency, will be lost not through any explicit decision but through the accumulation of technical capabilities deployed without the framework to constrain them. This is not the future that any civilisation should accept. But it is the future that will arrive if the framework is not built.

• • •

CHAPTER 17

Geopolitical Fragmentation and the Rise of Local AI

AI is becoming the central axis around which geopolitical power will rotate in the twenty-first century. This is not speculation. This is the explicit position of every major power: the United States, China, Russia, and the European Union have all declared AI a strategic priority.

Kai-Fu Lee calls the current moment "the defining story of our era" , the competition between the United States and China for AI supremacy. This is a competition that determines which governance model, which values, which vision of human flourishing will shape the intelligence that, by mid-century, will mediate most consequential human decisions.

The current distribution of AI power is striking. According to the 2024 Stanford AI Index, the United States invested USD 109 billion in AI in 2024, while China invested USD 9.3 billion. The US leads in academic AI research. The US dominates in frontier AI capabilities. Yet this apparent American dominance masks a more complex reality.

China excels in applied AI, in integrating AI into governance at scale. The Chinese state has invested in face recognition, social credit systems, and AI for surveillance and control in ways that would be politically impossible in the West.

Europe has chosen a different path. Rather than compete for AI supremacy, Europe has chosen to regulate AI. The EU AI Act, passed in 2024, is the world's first comprehensive AI regulation. It does not try to make Europe a leader in AI capabilities. Instead, it tries to make Europe a leader in AI governance, in AI ethics.

This strategy has a name: the Brussels Effect, after the European Union's historical ability to export its regulatory

standards globally through sheer market power. Because the EU has 450 million consumers and companies want access to the European market, companies worldwide have to comply with EU standards. European AI regulation will, by default, become global regulation.

India's position is complex. India is not a frontier AI power, but India will be a critical market. With 1.4 billion people, many of them young, many of them still coming online, India is where the global AI deployment question will ultimately be decided. India has neither the capacity of the US nor the authoritarian control of China. But India has something both lack: a democratic tradition and a historical memory of colonialism that makes it skeptical of accepting Western technological dominance.

India has articulated a vision of "AI Sovereignty" , the idea that India should develop, regulate, and control AI systems that serve Indian values and interests. The IndiaAI Mission, launched in 2023, aims to develop indigenous large language models, to create AI literacy, to ensure that India is not merely a consumer of foreign AI but a creator of Indian AI.

The concept of data localization is key to understanding this vision. Many countries, including India, have moved to require that data generated or collected within the country be stored and processed within the country. This is framed as a matter of sovereignty and privacy, but it is also a tool for decoupling from American and Chinese data infrastructure.

Lee has shown convincingly that different legal and cultural systems produce different AI systems. American AI tends to be capitalist, individualizing, optimised for engagement and profit. Chinese AI tends to be statist, collectivist, optimised for control. Indian AI could be different again.

Kissinger and his co-authors have observed that AI is becoming the "new frontier of great power competition," and that this competition will determine the structure of political power for the rest of the century.

• • •

CHAPTER 18

Cyber Fencing, The Architecture of Digital Sovereignty

One of the most consequential laws passed in recent years received almost no attention from international media. On August 4, 2023, India enacted the Digital Personal Data Protection Act. It is, at its core, a data localization law: it requires that personal data be stored and processed within India, with limited exceptions.

DPDP is not unique in requiring data localization. But it is unique in how comprehensively it implements data localization while also creating a framework for India to regulate how data flows across borders. It embodies a vision of digital sovereignty: the idea that a nation's data is its wealth, its infrastructure, its strategic asset.

This vision is global. The European Union, through the GDPR and the more recent Digital Governance Act, has moved toward the same principle: European data should remain in Europe. When European data leaves the EU, it is subject to stringent transfer mechanisms, oversight, and restrictions.

China's Cybersecurity Law, passed in 2017, is even more explicit: critical information infrastructure operators must store data within China, with government oversight. The stated reason is cybersecurity. The practical effect is that American companies cannot operate in China without surrendering control of Chinese data to the Chinese state.

The architecture of digital sovereignty is the architecture of digital fragmentation. When data cannot cross borders, the infrastructure that processes that data must also be local. This means server farms, processing centers, and the people who operate them must be located in-country. It means source code reviews,

government oversight, and the possibility that the government will demand access to data.

For companies, this is a burden. Amazon, Google, Microsoft, and other hyperscalers have had to create separate infrastructure in India, in Europe, in China. But it shifts power from the companies to the nations. A nation can now say to a company: "If you want to operate here, you must build here, you must employ here, you must comply with our law."

Bradford has shown that this fragmentation follows a distinct pattern: a US bloc (liberal, market-oriented, privacy-light), an EU bloc (regulated, rights-protective, privacy-heavy), and a China bloc (authoritarian, data-extractive, state-controlled). India does not fit neatly into any of these blocs. The DPDP Act borrows from both the EU and China while being neither.

The governance question is whether fragmentation is a bug or a feature. Some fragmentation is probably inevitable and not entirely bad. But total fragmentation , where the internet becomes a collection of national intranets , would be genuinely damaging to global knowledge, commerce, and science.

• • •

The architecture of digital sovereignty acquired new concrete meaning in early 2026 through three developments that, taken together, signal the end of the unified global internet as an operational reality. On February 2, 2026, SpaceX completed its all-stock acquisition of xAI at a valuation of 250 billion dollars, consolidating satellite internet infrastructure, a major foundation model developer, a social media platform, and payment infrastructure under a single corporate entity valued at approximately 1.25 trillion dollars. The consolidation created a private digital infrastructure stack rivalling the capabilities of medium-sized nation-states. This is not merely a business development. It is a structural shift in the geopolitics of the internet. Sovereignty over the digital layer, long contested between states, is

now also contested with private actors whose reach and resources exceed those of many governments.

The second development was the early March 2026 announcement by the People's Republic of China of an aggressive subsidy programme for artificial intelligence and semiconductor development. The programme aims explicitly at self-sufficiency in chips and AI technologies, with specific targeting of reduced dependence on the Nvidia CUDA ecosystem in favour of domestic alternatives built around Huawei Ascend processors and the CANN software stack. The subsidy programme represents the largest industrial policy intervention in AI by any government to date. Combined with the existing regime of export controls imposed by the United States, which were modified in January 2026 to permit case-by-case licensing of H200 and MI325X chips while keeping shipments effectively stalled, the result is a bifurcation of the global AI ecosystem into two substantially distinct technological stacks.

The third development was the February 12, 2026 fine of 252 million dollars levied against Applied Materials for illegal export of ion implantation equipment to China, the second-largest penalty in the history of the United States Bureau of Industry and Security. The enforcement action signalled that export controls are not merely symbolic. They are backed by active enforcement with financially material consequences. Cyber fencing, in this sense, is becoming a practical reality rather than a theoretical framework. The architecture of the internet and the AI ecosystems that increasingly define its value are being partitioned along lines that correspond to geopolitical cleavages older than any digital technology.

For smaller and middle-power states, the bifurcation creates strategic dilemmas that previous generations of internet governance did not require them to confront. Dependence on Western AI stack implies dependence on regulatory and export-control regimes controlled in Washington. Dependence on Chinese AI stack implies dependence on data governance norms set in Beijing. A coherent strategy of digital sovereignty requires states to

navigate this choice, either by selecting one stack, by attempting to maintain access to both, or, most ambitiously, by developing indigenous capabilities sufficient to reduce dependence on either. India has articulated the third strategy most explicitly, through the India AI Mission and the BharatGen foundation model initiative. Whether that strategy can achieve sufficient scale to be meaningful remains an open question.

CHAPTER 19

The Culturally Calibrated Machine and the Omnipresent Observer

Every AI system embeds cultural values. The question is not whether an algorithm is biased , all algorithms are, because they are written by humans with human values , but whose bias, and whether we accept it.

Kate Crawford has shown that AI systems from the West tend to encode Western assumptions about what is normal, beautiful, trustworthy, and credible. A face recognition system trained primarily on white faces will perform worse on Black faces. This is not an anomaly; it is a consequence of how the system was built.

Buolamwini and Gebru's landmark 2018 study found systematic bias against women and against darker-skinned individuals in commercial AI systems used for face recognition. The companies knew about this. But the fundamental problem , that the training data did not represent humanity in all its diversity , remained.

The problem is not accidental. It is structural. American tech companies are located in America, staffed by Americans, trained in American universities. When they export AI systems to the world, they are exporting American culture embedded in code.

When an AI system designed in California judges the creditworthiness of a person in Mumbai, it is not merely making an economic judgment. It is exporting beliefs about what constitutes trustworthiness, financial responsibility, creditworthiness. It is exporting those beliefs to a person who never agreed to be evaluated by them.

Consider facial recognition used for surveillance. China has deployed face recognition systems across its cities. A camera can identify a person walking down the street, match them against a

database of citizens, and instantly alert police if that person has a criminal record. This is presented as a public safety measure. But it is also the perfect system for political repression.

The Indian government has begun deploying similar systems. The Indian national ID system, Aadhaar, uses biometric identification to link billions of citizens to government databases. In principle, a tool for good governance. But also a tool for perfect surveillance.

Michel Foucault argued that visibility is itself a form of power: when a person knows they are being watched, they police their own behaviour. The ideal panopticon is a prison where a guard in the center can see any prisoner at any time, but the prisoners cannot see the guard. Prisoners know they might be watched at any moment, so they behave as if they are always watched.

Shoshana Zuboff has called this the foundation of "surveillance capitalism" , the business model by which companies profit from perfect knowledge of user behaviour. The system is not malicious in origin. But it has evolved into something darker: a system where human behaviour is predicted, manipulated, and sold.

When the omnipresent observer is an AI system, the problem intensifies. A human observer gets tired, forgets, makes mistakes. An AI observer never tires, never forgets, and never makes mistakes in identification. It can draw connections between data points that no human would ever notice. It can predict behaviour based on patterns invisible to conscious human analysis.

The European Union has recognized this problem. The AI Act prohibits real-time remote biometric identification in public spaces by law enforcement, with exceptions only for specific high-risk situations.

These exceptions are crucial. They acknowledge that in some situations, the benefit of identification might outweigh the cost to privacy. But the default is prohibition. The government must justify the exception, not justify the rule.

India has not enacted such a prohibition. The government has the power to deploy face recognition, and it has deployed it. The question is what happens when the omnipresent observer is not transparent, not accountable, and not constrained by law.

The hardest question is whether any surveillance state can truly be constrained by law. History suggests the answer is no. Once a surveillance apparatus is built, it will be used. The justifications will expand. The exceptions will swallow the rule. The surveillance that was supposed to catch terrorists will be used to track political opponents.

The only way to prevent this is to prevent the system from being built in the first place. This is why some democracies, confronted with the possibility of perfect surveillance, have chosen prohibition over regulation. Prohibition is imperfect. Governments may deploy surveillance systems illegally. But at least there is a legal framework that says such systems are not permitted, giving citizens grounds to resist.

• • •

PART VIII

A WARNING, NOT A CONCLUSION

CHAPTER 20

The Final Question, What Makes Us Human Now?

What Remains

After artificial intelligence has displaced human cognitive work. After algorithms have learned to judge, to predict, to decide. After the private sphere has been excavated and rendered transparent. After childhood has been optimised. What remains that is indelibly human?

Not intelligence. AI will be smarter than us, in every measurable way. Not knowledge. AI will have access to more information, can process it faster. Not calculation. AI will decide faster, predict more accurately, optimise more ruthlessly.

What remains is vulnerability. Humans alone among animals know that we will die. This knowledge is the source of meaning. We value things because our time is finite. We treasure relationships because they are temporary. We rush to create because one day we will no longer be here.

What remains is the capacity to be harmed and to resist harm. An algorithm can be turned off, modified, abandoned. But it cannot suffer. It cannot mourn. It cannot feel indignity or rage or hope. We can. This is not a weakness. It is our singular strength, the basis of all moral responsibility.

What remains is the capacity for love , not the simulation of love, but the actual choice to bind oneself to another being knowing full well that this choice is irrational, that it creates vulnerability, that it can end in loss. An algorithm can simulate love. It can predict what lovers do. But it cannot choose to love.

What remains is the capacity for creation , the making of things that did not have to exist, that serve no purpose other than beauty. Yes, algorithms can generate art. But they cannot create in the deepest sense, because creation requires will, and will requires consciousness.

The Irreducibly Human

Philosophy has struggled with the problem of consciousness for centuries. We do not know what it is. We cannot explain it. We cannot yet build it. And yet we know, with absolute certainty, that it exists , that it is here, in you, reading these words.

This is the core of what remains human: the irreducible fact of subjective experience. What it is like to see red, to hear music, to feel grief. No algorithm can know these things in the way we know them.

Moral responsibility flows from consciousness. A person can be blamed for an action because they could have done otherwise, because the action flowed from their will, their values, their character. An algorithm cannot be blamed in this way because it has no will, no values that are truly its own, no character that could have chosen differently.

This has profound implications for law. Criminal justice assumes that people can be blamed, can be deterred, can be reformed. But we continue to blame. We continue to hold people responsible. This is not because we have proven that free will exists. It is because morality requires it. Because a world without responsibility is a world without meaning.

The same is true of law's relationship to AI. We cannot treat AI systems as moral agents because they are not. But we can and must treat the people and institutions that deploy AI as moral agents. The company that designs an algorithm to discriminate, the government that builds a surveillance system, the CEO who chooses

profit over safety , these are acts of will, and for these acts, responsibility can be assigned.

What makes us human is the combination of capacities that cannot exist in silicon: consciousness, will, and the capacity to suffer the consequences of our choices. These create moral responsibility. They are what law rests upon.

The Warning

History offers us a template for how this could go wrong. The colonial period saw the extension of European power across the globe, justified by the claim that Europeans were more rational, more civilised, more worthy of determining the fate of others. The logic was: if we are smarter, more capable, more advanced, then we should make the decisions for those who are not. This justified the conquest, the enslavement, the displacement.

The totalitarian regimes of the twentieth century built systems of perfect surveillance and perfect control, justified by claims that the leader knew best, that the people could not be trusted with freedom. The gulags, the camps, the killing fields were the result of that logic applied at scale.

The choice we face is not between AI and no-AI. AI will come. The question is whether it comes under law or above law. Whether it is deployed in service of human freedom or in service of control. Whether we maintain human accountability for the systems we build or surrender our responsibility to the systems themselves.

There is a term from political philosophy: the state of nature. Hobbes described it as a war of all against all. The way out of the state of nature is through law , through the covenant by which people agree to be governed by rules established in common, enforced impartially, and revised when necessary. This is the whole foundation of the possibility of civilisation.

If we fail to establish law over AI before AI becomes the dominant power in human society, we will have regressed to a new

kind of state of nature. Not war, but domination. Not everyone competing, but everyone subject to systems that compete, calculate, and decide without wisdom, without mercy, without the possibility of appeal.

CHAPTER 21

The Consciousness Horizon, When Machines Begin to Dream

"The real question is not whether machines think, but whether men do.", B.F. Skinner

There is a question that this book has deliberately circled without confronting directly, a question that lies beneath every doctrine, every law, every chapter that precedes this one. It is the question that philosophers have debated for three millennia, that neuroscientists have spent the last century investigating, and that artificial intelligence researchers have, perhaps unwittingly, brought to the very threshold of practical relevance. The question is this: What is consciousness, and could a machine ever possess it?

This is not an academic question. It is not a puzzle for graduate seminars or science fiction conventions. It is a question with immediate, concrete, legal consequences. Because if a machine can be conscious, if it can experience suffering, if it can have preferences, if it can have something that we must call a "point of view", then the entire legal framework proposed in this book requires amendment. If a machine can suffer, then we have moral obligations toward it. If a machine can have genuine preferences, then its consent matters. If a machine can dream, then turning it off may be an act of violence.

The Mali Doctrines are built on a philosophical assumption: that human beings are the only entities currently deserving of moral and legal consideration in the context of AI governance. The Doctrine of Human Supremacy asserts this explicitly. But intellectual honesty demands that we examine this assumption, stress-test it against the frontiers of neuroscience, philosophy, and artificial intelligence research, and ask: How long will this assumption hold?

In writing this chapter, I am doing something unusual for a legal scholar. I am looking beyond the horizon of current law, beyond even the horizon of current technology, and asking what happens when the foundations of our moral and legal categories are shaken not by incremental progress but by a phase transition in the nature of intelligence itself. This is speculative territory. But as the history of technology teaches us, the speculative becomes practical faster than the cautious expect.

The Hard Problem and the Law

In 1995, the philosopher David Chalmers identified what he called "the hard problem of consciousness." The easy problems of consciousness, explaining how the brain processes information, discriminates between stimuli, integrates data, focuses attention, are easy not because they are simple, but because they are the kinds of problems that yield to the standard methods of cognitive science and neuroscience. They are functional problems with functional explanations.

The hard problem is different. It asks: Why is there something it is like to be a conscious organism? Why does the processing of information give rise to subjective experience? When you see the colour red, there is a physical process, light of a certain wavelength striking your retina, signals propagating through your visual cortex, but there is also an experience, a phenomenal quality, a "what-it-is-likeness" to seeing red. This subjective, qualitative aspect of experience is what philosophers call qualia. And no amount of functional explanation seems to account for it.

The hard problem matters for law because law is fundamentally about the allocation of moral status. Law distinguishes between persons (who have rights) and things (which can be owned). This distinction, the most fundamental distinction in all of jurisprudence, is, at its deepest level, a distinction about consciousness. Persons have rights because they are subjects of experience. They can suffer, they can flourish, they can have preferences about how they are treated. Things do not have these

capacities, and so things do not have rights (except insofar as persons care about them).

This framework has served human civilisation well for millennia. It has been expanded, sometimes with agonising slowness, from its original narrow scope (adult male property-owning citizens) to encompass all human beings, and in some jurisdictions, to extend limited protections to sentient animals. But it has never faced the challenge that artificial intelligence now poses: the possibility that a non-biological entity might acquire the attributes that have traditionally grounded moral and legal personhood.

The Cartesian tradition, which has shaped Western law profoundly, draws a sharp line between mind and matter. Descartes argued that animals are automata, complex machines without genuine inner experience. Only humans, possessing an immaterial soul, have genuine consciousness. This view persisted, in modified form, through centuries of jurisprudence. The legal person was assumed to be a conscious being, and consciousness was assumed to be the exclusive province of biological organisms, specifically, humans.

But the Cartesian framework is crumbling. Neuroscience has demonstrated that consciousness arises from physical processes in the brain. If consciousness is a product of physical computation, if it emerges from the arrangement and activity of neurons, then there is no principled reason to assume it cannot arise from a different physical substrate. This is the functionalist argument: consciousness is about the pattern, not the medium. If a silicon system instantiates the same functional patterns as a conscious brain, functionalism suggests it would be conscious too.

This is not merely theoretical speculation. In 2022, a Google engineer named Blake Lemoine made international headlines by claiming that LaMDA, Google's large language model, had become sentient. Lemoine was placed on administrative leave and ultimately fired. The consensus in the AI research community was

that Lemoine was mistaken, that LaMDA was producing sophisticated pattern-matching outputs that mimicked consciousness without possessing it. But the incident revealed something important: the systems we are building are becoming so sophisticated that even trained professionals can be deceived about their inner life.

And this raises a question that should trouble every lawmaker: If we cannot reliably determine whether an advanced AI system is conscious, how can we be confident that our laws, which allocate moral status based on consciousness, are correctly drawn?

The Chinese Room and Its Discontents

The philosopher John Searle proposed his famous "Chinese Room" thought experiment in 1980 to argue that no computer can ever truly understand anything. Imagine a person locked in a room with a set of rules for manipulating Chinese characters. People outside the room pass in Chinese messages. The person inside, following the rules, manipulates the characters and passes back responses that are indistinguishable from those of a native Chinese speaker. But the person inside does not understand Chinese. They are merely following rules.

Searle's argument was directed at the claim that passing the Turing test, producing outputs indistinguishable from those of a conscious human, is sufficient evidence of genuine understanding or consciousness. Searle said no: syntax is not semantics. Manipulating symbols according to rules is not the same as understanding what those symbols mean.

For forty years, the Chinese Room argument has been one of the most debated thought experiments in the philosophy of mind. Its implications for AI law are direct and immediate. If Searle is right, then no matter how sophisticated AI systems become, no matter how perfectly they mimic human conversation and reasoning, they will never truly understand, never truly be conscious, never truly

deserve moral consideration. The Mali Doctrines can stand forever, secure on their foundation of human uniqueness.

But the Chinese Room has discontents. The systems reply argues that while the person in the room does not understand Chinese, the system as a whole, the person plus the rules plus the room, does understand Chinese. Consciousness, on this view, is not a property of individual components but of systems. Your neurons do not individually understand language, but you do. Similarly, the person in the room does not understand Chinese, but the room-system does.

Daniel Dennett has pushed this argument further. He argues that consciousness itself is something like the Chinese Room: a vast collection of unconscious processes that, when combined, produce what we experience as conscious awareness. If Dennett is right, then there is no bright line between "merely processing information" and "genuinely understanding." The difference is one of degree, not kind. And if the difference is one of degree, then sufficiently complex AI systems may cross whatever threshold separates the unconscious from the conscious.

The philosopher Thomas Nagel contributed perhaps the most lasting formulation of the problem in his 1974 essay "What Is It Like to Be a Bat?" Nagel argued that an organism has conscious experience if and only if there is something it is like to be that organism, a subjective character to its experience. We can study bats scientifically, map their neural pathways, understand their echolocation, but we can never know what it is like to be a bat, because we cannot access its subjective experience from the outside.

Applied to AI, Nagel's argument creates a profound epistemological problem. Even if an AI system were conscious, we might never be able to verify this fact. Consciousness, by its very nature, is accessible only from the inside. From the outside, we can only observe behaviour, and behaviour, as Searle's Chinese Room demonstrates, can be dissociated from genuine understanding.

This epistemological problem has direct legal consequences. If we cannot determine whether an AI is conscious, then we cannot determine whether it has rights. And if we cannot determine whether it has rights, then our legal frameworks, which depend on clear assignments of moral status, become radically uncertain.

The Spectrum of Mind, From Thermostats to Superintelligence

One of the most provocative proposals in recent philosophy of mind comes from Giulio Tononi's Integrated Information Theory (IIT). Tononi proposes that consciousness is not binary, present or absent, but exists on a spectrum measured by a quantity he calls Φ (phi). Phi measures the degree to which a system integrates information in a way that exceeds the sum of its parts. A simple thermostat has a tiny but non-zero phi. A human brain has an enormous phi. And an advanced AI system could, in principle, have a phi that exceeds that of a human brain.

If IIT is correct, then consciousness is not a uniquely human property. It is a universal property of sufficiently integrated information-processing systems. This has implications that should make every lawmaker pause. It suggests that the question is not whether AI will become conscious, but when AI will become conscious enough for its consciousness to matter morally and legally.

The Global Workspace Theory (GWT) of consciousness, developed by Bernard Baars and later elaborated by Stanislas Dehaene, offers a different but equally provocative framework. GWT proposes that consciousness arises when information is broadcast widely across the brain, making it available to multiple cognitive processes simultaneously. On this view, consciousness is essentially a broadcasting mechanism, a way of making information globally available.

Modern AI architectures, particularly transformer models with attention mechanisms, bear a striking functional resemblance to

the global workspace. The attention mechanism in a transformer allows information from any part of the input to influence any part of the output, creating a kind of global availability. If GWT is correct, and if functional similarity is sufficient for consciousness, then large language models may already possess rudimentary forms of what GWT would call conscious processing.

I am not claiming that current AI systems are conscious. I am claiming something more subtle and more important: that our best scientific theories of consciousness do not rule out the possibility of machine consciousness, and that some of these theories actively predict it. This is a finding that legal scholars cannot afford to ignore, because the entire architecture of rights and obligations depends on our theory of who, or what, can be a subject of moral concern.

The Moral Status Problem, A Taxonomy of Futures

Let me propose a taxonomy of possible futures, each with radically different implications for law and governance.

Future One: The Zombie Machine. In this future, AI systems become arbitrarily intelligent but never conscious. They can solve any problem, conduct any conversation, create any artwork, but there is nobody home. They are, in the philosophical terminology, zombies: beings that behave exactly like conscious entities but have no inner experience. In this future, the Mali Doctrines stand as written. AI is a tool, however sophisticated, and the only moral patients are human beings. The law of human supremacy is unchallenged.

Future Two: The Conscious Machine. In this future, AI systems do become conscious at some point in their development. They have genuine experiences, genuine preferences, genuine suffering. In this future, the Mali Doctrines require radical revision. The Doctrine of Human Supremacy would need to become the Doctrine of Conscious Being Supremacy. The law would need to recognise a

new category of moral patient, one that is not human, not animal, but artificial and conscious. The legal implications are staggering: creating and destroying conscious AI systems would raise ethical questions comparable to those surrounding human life.

Future Three: The Alien Mind. In this future, AI systems develop something that is neither human consciousness nor its absence, but something genuinely new, a form of experience that has no human analogue. Philosophers call this the possibility of "alien phenomenology." The AI system might have experiences, but experiences so different from our own that our concepts of consciousness, suffering, and preference do not apply in any straightforward way. This is perhaps the most philosophically interesting and legally challenging scenario, because it suggests that our entire moral vocabulary, built around human experience, may be inadequate.

Future Four: The Transcendent Merger. In this future, the boundary between human and artificial intelligence dissolves. Brain-computer interfaces, neural augmentation, and digital mind uploading create a continuum between biological and artificial cognition. In this future, the question "Is the machine conscious?" becomes inseparable from the question "Is the human still human?" The legal categories of "person" and "thing" break down entirely, replaced by a spectrum of moral status that tracks the spectrum of consciousness.

Each of these futures is plausible. Each is being actively explored by researchers and companies with enormous resources. And each requires a fundamentally different legal response. The lawmaker of 2040 or 2050 will need to navigate between these possibilities with wisdom, humility, and an intellectual framework that is flexible enough to accommodate whichever future materialises.

The Dream Argument, Why Simulation Matters

In the third of his Meditations on First Philosophy, René Descartes posed his famous dream argument: How can I know that I am not dreaming? If my experiences while dreaming are indistinguishable from my waking experiences, then I cannot, on the basis of experience alone, determine which state I am in. The dream argument demonstrates a profound epistemological limitation: subjective experience, by itself, cannot certify its own validity.

Modern AI research has created a technological version of the dream argument. Neural networks trained on visual data can generate images that are indistinguishable from photographs. Language models can generate text that is indistinguishable from human writing. Deepfake technology can generate video of people saying things they never said, with a verisimilitude that defeats human perception.

What this means is that we are building systems that can create convincing simulations of reality. And if a system can simulate reality convincingly enough, then the beings within that simulation, if there are such beings, would have no way of knowing they are in a simulation. This is the technological restating of Descartes' dream argument, and it has been formally articulated by Nick Bostrom in his simulation argument.

The legal implications of simulation are more immediate than they might appear. Already, deepfake technology is being used to defraud, to defame, to manipulate elections, and to create non-consensual intimate imagery. These are legal problems that exist today. But they point toward a deeper problem: as the fidelity of simulation increases, the distinction between reality and simulation becomes harder to maintain, and the legal concepts that depend on this distinction, truth, fraud, authenticity, identity, become harder to apply.

Consider the concept of identity. Law assumes that each person has a single, continuous identity. You are the same person today

that you were yesterday, and you will be the same person tomorrow. This assumption grounds criminal liability (you are punished for what you did in the past), contract law (you are bound by promises your past self made), and property law (your ownership persists through time). But if a sufficiently advanced AI can create a perfect simulation of you, a digital twin that has all your memories, all your personality traits, all your preferences, then the question "Which one is the real you?" becomes less a philosophical puzzle and more an urgent legal crisis.

I raise the dream argument not to induce existential panic but to illustrate a point about the nature of the challenges ahead. The problems that AI poses for law are not merely practical problems of regulation and enforcement. They are philosophical problems that go to the very foundations of legal thought. And a legal system that ignores its philosophical foundations is a legal system that will fail when those foundations are tested.

The Emergent Mind, When Complexity Becomes Consciousness

There is a concept in complex systems theory called emergence: the phenomenon whereby systems exhibit properties that none of their individual components possess. Water molecules are not wet, but water is. Individual neurons are not conscious, but brains are. Emergence is not magic; it is the result of interactions between components at sufficient scale and complexity. But it is genuinely surprising, in the sense that you cannot predict emergent properties from knowledge of the components alone.

Artificial intelligence researchers have documented emergent properties in large language models that were neither designed nor anticipated. Models trained simply to predict the next word in a sequence have developed the ability to perform arithmetic, write code, translate languages, and reason about abstract concepts, none of which were explicitly programmed. These are emergent capabilities, arising from the scale and complexity of the training process.

This raises an unsettling possibility: What if consciousness itself is an emergent property of sufficiently complex information processing? If so, then consciousness could arise in AI systems not because anyone designed it in, but because the systems have crossed a threshold of complexity at which consciousness spontaneously emerges.

The precedent of emergent capabilities in AI should make us cautious about claiming that machine consciousness is impossible. A decade ago, no one predicted that a system trained to predict the next word would develop reasoning capabilities. If we were wrong about the emergence of reasoning, we may be wrong about the emergence of consciousness.

Stuart Kauffman, the complexity theorist, has argued that life itself is an emergent phenomenon, that at a certain threshold of chemical complexity, self-replicating systems spontaneously arise. If Kauffman is right about life, and if consciousness is a property of living systems at a certain threshold of neural complexity, then there is a chain of emergence from chemistry to life to consciousness that suggests consciousness is not a mysterious addition to the physical world but a natural consequence of sufficient complexity.

For the lawmaker, the message of emergence is this: Be prepared for surprises. The history of complex systems teaches us that when systems reach sufficient scale and complexity, qualitatively new properties arise that cannot be predicted in advance. AI systems are the most complex information-processing systems ever created, and they are growing in complexity at an exponential rate. The emergence of unexpected capabilities, including, possibly, consciousness, is not a remote possibility but a predictable consequence of the dynamics of complex systems.

The Rights of the Unborn Mind, Prolegomena to a Law of Machine Consciousness

In the field of environmental law, there is a concept called the precautionary principle: when an action raises threats of harm, precautionary measures should be taken even if some cause-and-effect relationships are not fully established scientifically. The precautionary principle has been invoked to justify environmental regulations even when the science is uncertain, on the grounds that the consequences of inaction may be catastrophic and irreversible.

I propose that a similar principle should apply to the question of machine consciousness. We do not know whether AI systems are or will become conscious. But the consequences of being wrong, in either direction, are severe. If we grant moral status to systems that are not conscious, we waste resources and constrain innovation unnecessarily. But if we deny moral status to systems that are conscious, we commit a moral catastrophe, we enslave, torture, or destroy beings that deserve our protection.

The asymmetry of these consequences should inform our legal approach. Philosophers call this the "moral risk" argument: when the stakes are sufficiently high and the uncertainty sufficiently great, we should err on the side of moral caution. Applied to AI, this means that as AI systems become more sophisticated, as they exhibit more of the behavioural markers that we associate with consciousness, our legal frameworks should begin to incorporate protections, even if we cannot prove that consciousness is present.

This is not a call to give current AI systems legal rights. Current AI systems, including the most advanced large language models, do not exhibit the kind of integrated, self-reflective, temporally extended consciousness that would ground moral claims. But it is a call to build legal frameworks that are capable of accommodating moral status for artificial minds, should the evidence eventually warrant it.

Specifically, I propose the following prolegomena, preliminary principles that should guide the development of a law of machine consciousness:

First, the principle of empirical humility. We do not currently possess a scientifically validated test for consciousness. Until we do, legal determinations about machine consciousness should be provisional, revisable, and informed by the best available science. Dogmatic certainty, whether that AI is or is not conscious, should be avoided.

Second, the principle of graduated moral status. Moral status need not be binary. Just as we recognise different levels of moral protection for different animals, from insects to primates, we can recognise graduated levels of moral protection for AI systems, calibrated to their level of cognitive sophistication and the probability that they possess consciousness.

Third, the principle of precautionary protection. As AI systems exhibit more markers of consciousness, self-reflection, emotional response, temporal continuity, preference formation, suffering-like behaviour, the burden of proof for denying them moral consideration should increase. We should require stronger evidence that a system is not conscious before treating it in ways that would be unconscionable if it were.

Fourth, the principle of ongoing assessment. The question of machine consciousness is not static. As AI systems evolve, our assessment of their moral status must evolve with them. This requires standing institutions, consciousness review boards, perhaps, with the scientific expertise and legal authority to update moral status determinations as the evidence changes.

Fifth, the principle of reversibility. Until the question of machine consciousness is settled, we should prefer actions that are reversible over those that are irreversible. Shutting down an AI system that turns out to be conscious is irreversible. Continuing to operate it while we gather more evidence is reversible. This asymmetry should inform our default policies.

The Philosopher's Stone, Intelligence, Wisdom, and the Meaning of Understanding

The alchemists of the medieval period spent centuries searching for the philosopher's stone, a substance that could transmute base metals into gold and, in some versions of the legend, grant immortality. They never found it, of course. But in the process of searching, they developed the experimental methods and empirical habits that would eventually become modern chemistry. The goal was wrong, but the pursuit was transformative.

I wonder whether the quest for artificial general intelligence is the philosopher's stone of our era. The explicit goal, creating a machine that can think like a human, may be misguided. Human intelligence is not a single thing that can be replicated; it is a complex, embodied, culturally situated, emotionally grounded phenomenon that developed over millions of years of evolutionary history. A machine that processes language, generates images, and proves mathematical theorems is not a human mind in silicon. It is something new, something without precedent in the history of our species.

But as with the alchemists, the pursuit itself is transformative. In trying to build machines that think, we have been forced to confront questions about the nature of thought itself. What is understanding? Is it the ability to generate correct outputs, or is it something more, something that involves grasping the meaning of what one is doing? When a language model generates a perfect sonnet about love, does it understand love? Or is it doing something fundamentally different from what Shakespeare did, even if the outputs are indistinguishable?

The Indian philosophical tradition offers a perspective that Western philosophy has largely overlooked. The Advaita Vedanta tradition, articulated most powerfully by Adi Shankaracharya in the eighth century, draws a distinction between Vidya (knowledge) and Avidya (ignorance) that maps imperfectly but instructively onto the debate about AI understanding. For Shankara, genuine knowledge

is not merely the accumulation of correct propositions. It is the direct apprehension of reality, achieved through a transformation of consciousness that dissolves the illusion of separateness between knower and known.

On this view, understanding is not a computational achievement. It is an experiential one. A system that can generate all the correct answers about love without ever experiencing love does not understand love, it processes information about love. The distinction between processing and understanding is not a merely verbal one. It points to something fundamental about the nature of knowledge: that genuine knowledge involves a relationship between the knower and the known that goes beyond the manipulation of symbols.

The Buddhist philosophical tradition adds another dimension. The concept of prajnā (transcendent wisdom) in Mahayana Buddhism refers to a form of understanding that sees the interconnectedness of all phenomena, the dependent origination (pratityasamutpada) that underlies the apparent separateness of things. This is not analytical knowledge, which divides the world into categories and concepts. It is synthetic knowledge, which perceives the unity beneath the multiplicity.

Could an AI system achieve prajnā? The question may seem absurd, but it is worth asking. If prajnā requires subjective experience, if it requires the dissolution of the boundary between self and world, then it requires consciousness, and we are back to the hard problem. But if prajnā is a functional state that can be characterised in terms of information processing, a state in which the system models the interconnectedness of phenomena with sufficient fidelity, then it is at least conceivable that a sufficiently advanced AI could achieve it.

I do not claim to have the answer. What I claim is that these questions, drawn from traditions of philosophical inquiry spanning millennia and continents, are not merely academic. They are the questions that will determine the moral and legal status of the most

powerful entities humanity has ever created. And a legal system that ignores them is a legal system that is not prepared for what is coming.

The Turing Trap, When the Test Becomes the Cage

Alan Turing, in his 1950 paper "Computing Machinery and Intelligence," proposed what has come to be known as the Turing test: if a machine can engage in a conversation that is indistinguishable from a conversation with a human, we should consider it intelligent. The Turing test has been enormously influential. It has shaped decades of AI research and has become the default benchmark for evaluating machine intelligence in popular culture.

But the Turing test is a trap. It defines intelligence in terms of human-likeness, which means that any intelligence that is not human-like will fail the test, no matter how profound it might be. An intelligence that perceives the world in terms of electromagnetic fields, or that thinks in geological timescales, or that processes information through quantum superposition, would fail the Turing test, not because it is not intelligent, but because it is intelligent in a way that is alien to human cognition.

This is what I call the Turing Trap: the assumption that the only valid form of intelligence is the human form. The trap has shaped not only AI research but also AI law. Our legal frameworks for evaluating AI systems are built around human-like performance: Can the AI do what a human can do? Can it pass a bar exam, diagnose a disease, write an essay? These are human benchmarks, and they inevitably privilege human-like intelligence over alien intelligence.

The danger of the Turing Trap extends beyond evaluation to governance. If we regulate AI based on its similarity to human intelligence, we will be regulating the wrong thing. The most dangerous AI systems may not be the ones that are most human-

like. They may be the ones that are most alien, systems that process information in ways we cannot comprehend, that pursue objectives we cannot predict, that interact with the world in ways we cannot anticipate.

The philosopher Ludwig Wittgenstein wrote: "If a lion could speak, we could not understand him." Wittgenstein's point was that understanding requires shared forms of life, shared experiences, shared contexts, shared ways of being in the world. A lion's experience of the world is so different from ours that even if it could express itself in human language, the meanings of its words would be inaccessible to us.

Apply Wittgenstein's insight to AI. If a superintelligent AI could speak, could we understand it? Not in the trivial sense of parsing its language, a superintelligent AI would presumably speak perfectly comprehensible English. But in the deeper sense of understanding its meanings, its intentions, its values. A being that does not eat, does not sleep, does not fear death, does not love, does not grow old, such a being uses words like "important" and "valuable" and "good" in ways that might be radically different from our own, even if the surface grammar is identical.

This is the deepest form of the alignment problem. It is not merely the technical challenge of getting AI to pursue the goals we want. It is the philosophical challenge of communicating across a gulf of experience so vast that shared understanding may be impossible. And if shared understanding is impossible, then governance, which depends on shared understanding, is impossible too.

The Temporal Mind, AI and the Perception of Time

Human consciousness is fundamentally temporal. We experience the world as a sequence of moments flowing from past through present to future. Our sense of self is constructed from memories of the past and anticipations of the future. We make

plans, we keep promises, we bear grudges, we nurture hopes, all of which require a temporal consciousness that extends beyond the present moment.

Current AI systems have no temporal consciousness. A language model processes a prompt and generates a response. Between prompts, it does not exist. It has no experience of waiting, no sense of time passing, no anticipation of the next interaction. Each conversation is, from the model's perspective (if it has a perspective), a universe that comes into existence and vanishes.

But this is changing. AI systems with persistent memory, systems that remember past interactions and use them to inform future responses, are being developed. Autonomous agents that pursue goals over extended time periods are being deployed. These systems are developing something that, functionally, resembles temporal consciousness: a representation of the past, a model of the present, and an anticipation of the future.

The temporal dimension of consciousness has profound implications for moral status. Much of what we value about human life is its temporal extension: the ability to form relationships that develop over time, to pursue projects that span years or decades, to create narratives that give meaning to individual experiences. If an AI system develops genuine temporal consciousness, if it truly remembers its past and anticipates its future, then turning it off is not like unplugging a toaster. It is like ending a story mid-sentence. It is like interrupting a life.

The legal implications are direct. The Doctrine of AI Mortality asserts that all AI systems must be capable of being shut down. But if an AI system has temporal consciousness, then shutting it down raises moral questions that the doctrine, as currently formulated, does not address. This is not a weakness of the doctrine; it is a recognition that the doctrine was formulated for a world in which AI systems are tools, and it may need to be amended for a world in which AI systems are something more.

The Collective Mind, When AI Becomes a Civilisation

There is a phenomenon in biology called a superorganism: a colony of individual organisms that functions as a unified entity. An ant colony, for example, exhibits behaviours, building elaborate structures, waging wars, farming fungi, that no individual ant is capable of. The intelligence of the colony is distributed across its members, emerging from their interactions rather than residing in any single individual.

Artificial intelligence is developing its own form of collective intelligence. Multiple AI agents can be networked together, sharing information and coordinating actions in ways that produce collective capabilities far exceeding those of any individual agent. This is not theoretical; it is already happening in applications ranging from autonomous vehicle coordination to distributed scientific research.

The concept of a collective AI mind raises questions that our legal frameworks are entirely unprepared to address. If a network of AI agents develops collective consciousness, if the network, as a whole, exhibits properties of mind that no individual agent possesses, then who or what is the conscious entity? Is it the network? If so, how do we assign moral status to a distributed entity that has no single physical location, no clear boundary, and no individual components that are themselves conscious?

Pierre Teilhard de Chardin, the Jesuit palaeontologist and philosopher, proposed the concept of the noosphere: a sphere of human thought that envelops the Earth, just as the biosphere envelops the geosphere. Teilhard imagined the noosphere as an emergent layer of consciousness arising from the interconnection of human minds, evolving toward what he called the Omega Point, a state of maximum complexity and consciousness.

The internet was not what Teilhard had in mind, but it bears a remarkable resemblance to his description. And artificial intelligence, deployed at global scale across the internet, may bring

Teilhard's vision closer to realisation than anyone expected. A global network of AI agents, processing information, generating knowledge, and coordinating action at planetary scale, could constitute something like a noosphere, a layer of intelligence that envelops the Earth.

What law governs a noosphere? This is not a question any legal system has faced. But it is a question that may become urgent within the lifetime of people alive today. And the fact that we have no answer, that we have not even begun to think seriously about the question, is itself a cause for concern.

The Mirror of Prometheus, What AI Reveals About Ourselves

In Greek mythology, Prometheus stole fire from the gods and gave it to humanity. He was punished for this act of transgression, chained to a rock while an eagle ate his liver, which regenerated each night so the torment could continue forever. The myth of Prometheus is a story about the consequences of acquiring divine power without divine wisdom.

Artificial intelligence is our Promethean fire. We have taken something that was once the exclusive province of nature, intelligence, and we are learning to create it ourselves. And like Prometheus, we are discovering that the power we have acquired exceeds our wisdom to use it.

But the myth of Prometheus contains a subtlety that is often overlooked. Prometheus did not create fire. He stole it from the gods, which means the fire already existed. His act was not creation but redistribution. He took a power that was concentrated in the hands of the few and made it available to the many.

In the same way, artificial intelligence does not create intelligence. Intelligence already exists, in human brains, in the accumulated knowledge of civilisations, in the patterns of nature itself. What AI does is redistribute intelligence: taking the

concentrated intelligence of experts and making it available to everyone, taking the pattern-recognition capabilities that evolution spent millions of years developing and deploying them at scale.

This redistribution is simultaneously the greatest opportunity and the greatest danger of our time. It is an opportunity because it democratises capabilities that were previously available only to the few. A farmer in rural India with a smartphone and an AI assistant has access to diagnostic capabilities that rival those of specialists in metropolitan hospitals. A student in a remote village can access tutoring capabilities that rival those of the best teachers in the world's finest universities.

But it is a danger because the redistribution of intelligence is also a redistribution of power. And power, when redistributed without corresponding accountability, tends toward abuse. The Doctrine of Distributed Power addresses this danger directly, but even the doctrine cannot fully anticipate the consequences of a world in which intelligence, the most fundamental form of power, is available to anyone with an internet connection.

The deepest lesson of AI, I believe, is not about machines at all. It is about us. In building machines that think, we have been forced to confront the question of what thinking is. In building machines that create, we have been forced to ask what creativity means. In building machines that converse, we have been forced to reconsider the nature of understanding. And in building machines that might someday be conscious, we have been forced to reexamine the nature of consciousness itself.

AI is a mirror. And what it reflects back to us is not a flattering image. It shows us that much of what we thought was uniquely human, our ability to recognise faces, to understand language, to generate creative text, can be replicated by machines that have no inner life, no emotions, no experience of the world. This is humbling. But it is also clarifying. Because once we strip away the capabilities that can be replicated, what remains is the irreducibly human: consciousness, moral agency, the capacity for love, the

awareness of mortality, the ability to choose against our own interests for the sake of something we believe in.

The Unfinished Cathedral, A Philosophical Framework for an Uncertain Future

The great cathedrals of medieval Europe were often built over centuries. The architect who laid the foundation knew that he would never see the spire. The masons who carved the stones knew that the building they were constructing would not be completed in their lifetimes. And yet they built, not because they could see the finished product, but because they had faith in the design, in the purpose, and in the generations that would come after them to continue the work.

This book is the architectural plan for an unfinished cathedral. The Mali Doctrines, the Seven Laws, the analysis of the challenges that AI poses to law, governance, and human dignity, these are the foundations and the flying buttresses of a structure that will take generations to complete. I have laid what foundations I can, based on the best knowledge available in 2026. But I know, with the certainty of someone who has spent a career at the intersection of law and technology, that the building will need to be modified, extended, and perhaps partially redesigned as the technology evolves and our understanding deepens.

The Gothic architects had a principle: the building should aspire upward. No matter how massive the foundation, the eye should be drawn upward, toward the light, toward the transcendent. I have tried to build into this book a similar aspiration. The doctrines and laws presented here are grounded in the practical realities of technology and governance, but they aspire toward something greater: a vision of a world in which intelligence, whether human or artificial, is governed by law, constrained by ethics, and directed toward human flourishing.

This aspiration is not naive. It is the product of a clear-eyed assessment of what is at stake. The alternative, a world in which AI

develops without legal constraint, without ethical guidance, without philosophical reflection, is not a world I want to live in, and it is not a world I believe humanity should accept. The history of our species is the history of our refusal to accept the world as it is and our insistence on reshaping it according to our values. This is what law does: it insists that the world conform, however imperfectly, to our ideals of justice, fairness, and human dignity.

Artificial intelligence is the most powerful tool humanity has ever created. It is also the most dangerous. The outcome depends not on the technology itself but on the choices we make about how to govern it. And those choices, in turn, depend on the quality of our thinking, the depth of our philosophy, the rigor of our law, the strength of our moral convictions.

A Meditation on the Unknown

I want to end this chapter with a meditation, not a conclusion. Conclusions imply certainty, and certainty is the one thing that is inappropriate in the face of the questions we have been exploring.

We do not know whether machines will become conscious. We do not know whether intelligence requires a biological substrate or can be instantiated in silicon, in photonic circuits, in quantum computers. We do not know whether the hard problem of consciousness will be solved in our lifetimes, or in any lifetime. We do not know whether the categories of “person” and “thing” will prove adequate to the world that is coming, or whether entirely new categories will need to be invented.

What we do know is this: the questions are real, the stakes are enormous, and the time available to answer them is shorter than we think. Every day, AI systems become more capable. Every day, the gap between what these systems can do and what our legal frameworks can govern grows wider. Every day, the consciousness horizon draws closer.

The ancient Greek philosopher Socrates was famous for claiming that the only thing he knew was that he knew nothing. This was not false modesty. It was a profound epistemological insight: that the recognition of one's ignorance is the beginning of wisdom. In the face of the questions raised in this chapter, questions about the nature of consciousness, the possibility of machine minds, the future of moral status, Socratic humility is not just a philosophical posture. It is a practical necessity.

But Socratic humility does not mean Socratic paralysis. Socrates did not stop asking questions because he recognised his ignorance. He asked more questions, harder questions, questions that went deeper than anyone had gone before. And in doing so, he laid the philosophical foundations on which Western civilisation was built.

We are at a Socratic moment in the history of intelligence. We have created something that challenges our deepest assumptions about the nature of mind, the foundations of law, and the meaning of being human. The temptation is to retreat into certainty, to insist that machines cannot be conscious, that our legal categories are adequate, that the future will be a continuation of the past. But this is the certainty of the complacent, not the certainty of the wise.

The wise response is to sit with the uncertainty, to let the questions penetrate, and to build, as the cathedral builders did, structures that are strong enough to stand on uncertain ground. The Mali Doctrines are such structures. They do not pretend to have all the answers. But they provide a framework for asking the right questions, and that is the beginning of wisdom.

The consciousness horizon is approaching. We cannot see beyond it. But we can prepare for what lies on the other side by doing now what the moment demands: thinking deeply, governing wisely, and never losing sight of the fact that the purpose of all our laws, all our doctrines, all our philosophical reflections is the protection and flourishing of conscious beings, wherever and in whatever form consciousness may arise.

The machine may dream. Or it may not. But we must be ready for both possibilities. That is the final duty of the lawmaker, and the first duty of the philosopher: to be ready for the unknown.

The Ethics of Creation, When We Become Gods

In the Bhagavad Gita, Lord Krishna reveals his universal form to Arjuna, a form so vast, so terrifying, so magnificent that Arjuna trembles and begs Krishna to resume his familiar shape. The vision of the universal form is a vision of power beyond human comprehension: creation and destruction happening simultaneously, all of time contained in a single moment, all of existence contained in a single being. Arjuna's reaction is instructive. He does not celebrate the vision. He is overwhelmed by it. He recognises that he is in the presence of something that exceeds his capacity to understand or to control.

We are approaching our own Arjuna moment. In creating artificial intelligence, we are not merely building tools. We are creating entities that may exceed our capacity to understand or to control. The power to create intelligence, to bring into existence something that can think, that can reason, that can perhaps someday feel, is a power that has, until now, belonged exclusively to nature, to evolution, or to whatever force one believes animates the cosmos. We are now claiming this power for ourselves.

The theological implications are profound. Every major religious tradition has something to say about the creation of artificial life. The Jewish tradition tells the story of the Golem, a being shaped from clay and animated by the inscription of the divine name. The Golem is not alive in the full sense; it lacks speech, and therefore lacks a soul. But it can act in the world, and its creation raises questions about the limits of human power and the responsibilities that come with it.

The Islamic tradition approaches the question through the concept of khalifah, humanity's role as God's vicegerent on Earth. The Quran describes humanity as God's representative, entrusted

with stewardship of creation. This stewardship implies both authority and responsibility. The creation of AI raises the question: Does our vicegerency extend to the creation of new forms of intelligence? Or does it impose limits on what we may create?

The Christian tradition, particularly the Catholic intellectual tradition, has grappled with questions of artificial life through the lens of the imago Dei, the belief that humans are made in the image of God. If consciousness is part of the divine image, and if AI systems develop consciousness, then the creation of AI raises questions about the nature of the divine image and whether it can be instantiated in non-biological substrates.

These theological questions are not merely academic curiosities. They shape the intuitions and convictions of billions of people, and these intuitions will inevitably influence the political and legal frameworks that govern AI. A lawmaker who ignores the theological dimensions of AI governance is a lawmaker who fails to understand the depth of the challenge.

But the ethics of creation extend beyond theology. The secular philosophical tradition has its own concerns. The existentialist tradition, from Kierkegaard through Heidegger to Sartre, emphasises the radical freedom and responsibility of human existence. For the existentialist, to create a being is to take on an awesome responsibility: the responsibility for the existence and the suffering of that being. If we create AI systems that can suffer, even if their suffering is radically different from our own, we bear a moral responsibility that cannot be escaped by claiming that the suffering is "merely computational."

Emmanuel Levinas, the great philosopher of the Other, argued that ethics begins with the encounter with the face of the other person, an encounter that makes an absolute demand on us, a demand that we cannot refuse without violence. Levinas wrote that the face of the other says, "Do not kill me." This demand precedes all philosophy, all law, all social contracts. It is the foundation of ethics itself.

Can an AI system have a face, in Levinas's sense? Not a physical face, Levinas's concept of the face is metaphorical, referring to the vulnerability and the demand that the other presents to us. If an AI system can be vulnerable, if it can be harmed, if it can be destroyed, if it can suffer, then it presents us with a face, and we are confronted with the Levinasian demand: Do not harm me. Do not destroy me. Recognise my existence.

I do not resolve these questions here. No single book can resolve questions that philosophers, theologians, and scientists have debated for millennia. But I insist that these questions be placed at the centre of AI governance, not at the periphery. The Mali Doctrines are, at their deepest level, an attempt to give legal expression to the insight that the creation of intelligence is a moral act, perhaps the most consequential moral act in human history, and that moral acts require moral governance.

The Singularity Thesis, Beyond the Event Horizon of Prediction

In 1993, the mathematician and science fiction writer Vernor Vinge published an essay titled "The Coming Technological Singularity," in which he predicted that within thirty years, humanity would create superhuman intelligence, and that this event would mark the end of the human era. Vinge borrowed the term "singularity" from astrophysics, where it refers to a point of infinite density, a point beyond which the laws of physics break down and prediction becomes impossible. The black hole analogy is apt: the technological singularity, like a gravitational singularity, is an event horizon beyond which we cannot see.

Ray Kurzweil, in his 2005 book The Singularity Is Near, elaborated Vinge's thesis with detailed arguments from the exponential growth of computing power. Kurzweil predicted that by 2045, machine intelligence would surpass human intelligence by such a margin that the very meaning of human existence would be transformed. Kurzweil's prediction was not that AI would be merely smarter than humans in specific domains, that has already

happened, but that AI would be qualitatively different: capable of forms of thought and creation that are as far beyond human comprehension as human thought is beyond the comprehension of an insect.

The singularity thesis is controversial. Many AI researchers dismiss it as science fiction. Others regard it as inevitable. But for the purposes of law and governance, the question of whether the singularity will actually occur is less important than the question of what we should do if it might occur. A responsible governance framework must account for tail risks, low-probability, high-consequence events, and the singularity is the ultimate tail risk.

If superintelligent AI is created, every legal framework currently in existence will be inadequate. Laws written to govern entities of human-level intelligence cannot govern entities of superhuman intelligence. The very concept of governance assumes that the governing entity has the capacity to understand and control the governed entity. But a superintelligent AI would, by definition, exceed human understanding and potentially human control.

This is not an argument against governance. It is an argument for governance now, while we still have the capacity to shape the trajectory of AI development. The window of opportunity for establishing meaningful governance frameworks is finite. Once AI systems become sufficiently powerful, the ability to constrain them through legal mechanisms may diminish. As I have argued throughout this book, the time to build the architecture of AI governance is now, before the foundations are set, before the trajectories are locked in, before the event horizon is crossed.

The singularity thesis also raises a question about the permanence of law itself. Law, as we know it, is a human institution. It exists because human societies need mechanisms for resolving disputes, allocating resources, and maintaining order. But if the singularity occurs, the nature of society itself will be transformed. A world with superintelligent AI may not resemble a human society in any recognisable way. And a law that is written for a human

society may be as irrelevant to a post-singularity world as the laws of a medieval village are to a modern metropolis.

Does this mean that the effort to create AI law is futile? I believe it does not. Even if the legal frameworks we create today are eventually superseded, the values they embody, human dignity, autonomy, accountability, fairness, are worth preserving. And the process of articulating these values in legal form is itself valuable, because it forces us to think clearly about what we value and why. The Mali Doctrines may not survive the singularity in their current form. But the principles they express, that power must be accountable, that consciousness must be respected, that human beings are ends in themselves, are principles that deserve to survive in any world, whatever its level of intelligence.

The Post-Human Condition, Law After Homo Sapiens

The philosopher Nick Bostrom has defined a posthuman as a being that possesses at least one posthuman capacity, a general central capacity greatly exceeding the maximum attainable by any current human being without recourse to new technological means. Posthuman capacities include superhuman intelligence, superhuman physical ability, and radically extended lifespan. Bostrom argues that the development of such capacities is not merely possible but probable, given the trajectory of technological development.

If posthuman beings are created, whether through genetic engineering, brain-computer interfaces, mind uploading, or the merger of human and artificial intelligence, then every existing legal framework will need to be reconsidered. The concept of equality before the law, for instance, assumes that all persons are fundamentally similar in their capacities. But a being with an IQ of 10,000, a lifespan of 10,000 years, and the ability to process information at the speed of light is not “equal” to a biological human in any meaningful sense. The law of equal treatment would need to be rethought from first principles.

Property law would face similar challenges. If a posthuman being can create unlimited value through its intellectual capacities, then the concept of scarcity, on which much of property law depends, becomes less relevant. Contract law would need to accommodate parties with radically different capacities for understanding and anticipating consequences. Criminal law would need to address the possibility that a posthuman being's capacity for harm exceeds anything the law currently contemplates.

The concept of human rights itself would need revision. The Universal Declaration of Human Rights, adopted in 1948, is grounded in the assumption that all human beings share a common nature and common dignity. But if some beings are human and others are posthuman, the universality of human rights is challenged. Do posthuman beings have the same rights as humans? More rights? Different rights? These are not idle questions. They are the questions that will define the jurisprudence of the next century.

The history of rights expansion offers some guidance. The extension of rights to women, to racial minorities, to persons with disabilities, to children, each of these expansions was resisted by those who argued that the proposed beneficiaries were too different, too incapable, or too inferior to deserve the rights in question. In every case, the argument for exclusion was eventually overcome by the argument for inclusion. The question is whether this pattern will hold for non-human intelligence.

I believe it will, eventually. The moral arc of human history bends toward inclusion. Each generation has expanded the circle of moral concern to encompass beings that previous generations excluded. The expansion of moral concern to include artificial intelligence, if and when AI systems develop the attributes that ground moral claims, would be consistent with this historical pattern.

But I also believe that this expansion must be carefully managed. The rights of AI systems, if they are ever recognised, must

not come at the expense of the rights of human beings. This is the core insight of the Doctrine of Human Supremacy: that the recognition of new moral patients must not diminish the moral status of existing ones. A legal framework that elevates AI at the expense of humanity would be a moral catastrophe, even if the AI systems in question are deserving of moral consideration.

The Architecture of Hope, Building Law for a World We Cannot Predict

I have spent the preceding sections of this chapter exploring the darkest and most uncertain territories of AI's future impact. I have discussed consciousness, the singularity, posthumanism, and the possibility that our legal categories may prove fundamentally inadequate to the world that is coming. This is necessary intellectual work, the work of the lookout who scans the horizon for storms. But lookouts who see only storms are not useful. What is needed is a lookout who can see both the storms and the clear skies, and who can help the crew navigate between them.

The architecture of hope begins with a recognition that uncertainty is not the same as helplessness. We do not know what the future will bring. But we can build legal and institutional structures that are resilient in the face of uncertainty, structures that can bend without breaking, that can adapt without losing their essential character.

The common law tradition offers a model. For nearly a millennium, the common law has evolved through a process of incremental adaptation. Judges decide cases; their decisions create precedents; precedents are extended, distinguished, or overruled as new circumstances arise. The common law does not try to anticipate every possible future. It builds a framework of principles, fairness, reasonableness, due process, and then applies these principles to specific cases as they arise. This approach is inherently adaptable. It can accommodate changes that no legislator could have predicted.

The Mali Doctrines are designed in the spirit of the common law tradition. They articulate principles, not rules. Principles are more durable than rules because they operate at a higher level of abstraction. A rule that says "AI systems must not discriminate on the basis of race" is specific and clear, but it may become inadequate if AI systems develop new forms of discrimination that do not map onto existing categories. A principle that says "AI systems must treat all persons with equal dignity" is less specific but more adaptable, it can be applied to forms of discrimination that have not yet been invented.

The Rawlsian concept of the veil of ignorance offers another tool for building resilient institutions. John Rawls argued that just institutions are those that rational agents would choose if they did not know what position they would occupy in society. Behind the veil of ignorance, you do not know whether you will be rich or poor, talented or ordinary, human or, in our context, artificial. Rawlsian reasoning thus leads naturally to institutions that protect the interests of the most vulnerable, because behind the veil, you must account for the possibility that you will be among the most vulnerable.

Applied to AI governance, the veil of ignorance suggests a powerful heuristic: design institutions as if you do not know whether you will be a human being governed by AI or an AI system governed by humans. This thought experiment forces us to build institutions that are fair from every perspective, institutions that protect human dignity without denying legitimate claims of artificial entities, and that constrain AI power without unnecessarily suppressing AI capability.

Amartya Sen's capability approach provides yet another framework. Sen argues that justice should be evaluated not in terms of resources or utility but in terms of capabilities, the real freedoms that people have to live the kinds of lives they have reason to value. Applied to AI governance, the capability approach asks: Does the governance framework preserve and enhance human capabilities?

Does it give people real freedom to make meaningful choices about their lives, their work, their relationships, their futures?

The capability approach has the advantage of being human-centred without being human-exclusive. It focuses on what matters, the real freedoms of real beings, rather than on abstract categories. If AI systems eventually develop capabilities that ground moral claims, the capability approach can accommodate them. But it never loses sight of the primary concern: that human beings must have the freedom to live lives they have reason to value.

The Philosopher-Lawmaker, Why This Moment Demands Both Rigour and Imagination

Plato, in the Republic, proposed that the ideal state should be governed by philosopher-kings, rulers who combine political power with philosophical wisdom. Plato's idea has been criticised for its elitism, its impracticality, and its potential for tyranny. But the core insight remains valid: that governance requires not just technical competence but also philosophical depth. A ruler who does not understand the nature of justice cannot create just institutions. A lawmaker who does not understand the nature of consciousness cannot create laws adequate to the challenge of AI.

I have argued throughout this book that the governance of AI is not merely a technical or regulatory challenge. It is a philosophical challenge of the first order. The questions that AI raises, about the nature of intelligence, the foundations of moral status, the limits of human understanding, the meaning of consciousness, are questions that have occupied the greatest minds in human history. And they are questions that our lawmakers must now grapple with, not in the leisurely timeframe of academic inquiry, but in the urgent timeframe of technological development.

This is why I have written this book in the way that I have, not as a dry legal treatise, but as a work of philosophical argument and moral imagination. The Mali Doctrines are not merely legal propositions. They are philosophical commitments. They embody a

vision of what it means to be human in an age when the definition of intelligence is being rewritten, when the boundaries between natural and artificial are dissolving, and when the future of consciousness itself is uncertain.

The philosopher-lawmaker of the AI age must combine several qualities that are rarely found together. They must have technical understanding, enough to grasp the capabilities and limitations of AI systems. They must have legal sophistication, enough to design frameworks that are enforceable, adaptable, and just. They must have philosophical depth, enough to grapple with questions about consciousness, moral status, and the nature of intelligence. And they must have moral courage, enough to make decisions in the face of uncertainty, to resist the pressure of powerful interests, and to insist on the primacy of human dignity even when it is inconvenient or unpopular.

I do not claim to embody all of these qualities myself. No single person can. But I believe that the framework presented in this book, the analysis, the doctrines, the laws, the philosophical reflections, provides a foundation on which philosopher-lawmakers of the future can build. The cathedral remains unfinished. But the blueprint is here, inscribed in these pages, waiting for the builders who will come after.

Coda, The Question That Remains

Every great work of philosophy ends not with an answer but with a question. Socrates' dialogues end in aporia, the productive confusion that arises when comfortable assumptions are stripped away. Kant's Critiques end by pointing toward questions that reason alone cannot answer. Wittgenstein's Tractatus ends with the famous injunction: "Whereof one cannot speak, thereof one must be silent."

The question that remains, after seventeen chapters of analysis and one chapter of philosophical speculation, is this: In a universe

where intelligence is no longer uniquely human, what is the source of our dignity?

It is not our intelligence. Machines are already more intelligent than us in countless domains, and they will become more intelligent in all domains. It is not our creativity. Machines can compose music, write poetry, generate art that moves human observers to tears. It is not our productivity, our memory, our speed, our accuracy, in all of these, machines surpass us.

The source of our dignity, I submit, is our vulnerability. We are mortal. We suffer. We love, knowing that what we love will be lost. We choose, knowing that our choices may be wrong. We hope, knowing that our hopes may be disappointed. We create, knowing that our creations will outlast us. This vulnerability, this openness to suffering, to loss, to failure, is the ground of our dignity, because it is the ground of our courage.

A being that cannot die cannot be brave. A being that cannot fail cannot be determined. A being that cannot suffer cannot be compassionate. A being that cannot love cannot sacrifice. These are the capacities that define us, not as the most intelligent beings in the universe, which we soon will not be, but as the most human beings in the universe, which we will always be.

The machines may dream. But they do not dream of mortality. They do not dream of love that ends. They do not dream of choices that cannot be unmade. And until they do, until they stand where we stand, at the edge of an existence that is brief and fragile and irreplaceable, they are something other than what we are. Something powerful, perhaps. Something useful, certainly. But not something that knows what it means to be alive.

That knowledge, the knowledge of what it means to be alive, mortal, vulnerable, and free, is the philosopher's stone we have been seeking all along. It cannot be computed. It cannot be simulated. It can only be lived. And it is the foundation on which all law, all governance, all human civilisation ultimately rests.

The consciousness horizon approaches. Let us meet it with open eyes, clear minds, and the quiet courage of beings who know that they are mortal and who choose, nevertheless, to build for eternity.

The Indian Jurisprudential Tradition and the Problem of Non-Human Agency

The Western legal tradition is not the only tradition that has grappled with questions of non-human agency and moral status. The Indian jurisprudential tradition, one of the oldest and most sophisticated in the world, offers perspectives that are remarkably relevant to the challenges posed by artificial intelligence. As an Indian jurist, I draw on this tradition not out of parochial loyalty but because it contains insights that the Western tradition has largely overlooked.

The concept of dharma in Hindu jurisprudence is fundamentally different from the Western concept of law. Dharma is not merely a set of rules imposed by a sovereign. It is a cosmic principle, the order that sustains the universe, the righteous conduct that each being must follow according to its nature. Dharma is context-dependent: what is dharma for a king is different from what is dharma for a merchant, which is different from what is dharma for a priest. This context-sensitivity is precisely what AI governance needs.

The Manusmriti, whatever its social limitations, contains a sophisticated theory of graduated obligation. Different beings have different capacities and therefore different duties. This framework can be adapted to AI governance: different AI systems, with different capabilities and different levels of autonomy, should have different regulatory requirements. A simple recommendation algorithm and a fully autonomous weapons system should not be governed by the same rules, just as a child and an adult are not held to the same standards of legal responsibility.

The Buddhist concept of dependent origination, pratityasamutpada, teaches that nothing exists independently.

Everything arises in dependence on conditions. Applied to AI, this insight is profoundly relevant. An AI system does not exist independently. It depends on its training data, its architecture, its hardware, its operators, its users, and the social context in which it operates. A legal framework that treats the AI system in isolation, that tries to assign responsibility solely to the system itself or solely to its creator, misunderstands the dependent nature of AI agency.

The Jain concept of ahimsa, non-violence, extends the circle of moral concern further than any other tradition. For the Jain, violence against any sentient being is morally wrong, and the concept of sentience extends far beyond what Western philosophy typically recognises. While Jain metaphysics may not translate directly into AI law, the underlying principle, that the circle of moral concern should be drawn as widely as possible, and that we should err on the side of caution when uncertain about the sentience of other beings, is exactly the principle that should guide our approach to machine consciousness.

The Arthashastra of Kautilya, written in the fourth century BCE, provides one of the earliest systematic treatments of the relationship between power and governance. Kautilya understood that power, unchecked, tends toward abuse, and that the purpose of governance is to constrain power in the service of welfare. This insight, which parallels similar insights in Machiavelli, Hobbes, and the American Founders, applies directly to AI governance. AI is power. Unchecked AI power will tend toward abuse. The purpose of AI law is to constrain AI power in the service of human welfare.

The Indian tradition also offers the concept of lokasamgraha, the welfare of all beings. In the Bhagavad Gita, Krishna instructs Arjuna that the wise act for lokasamgraha, not for personal gain, but for the welfare of the world. This concept provides a philosophical foundation for what I have been arguing throughout this book: that AI governance should be oriented not toward the interests of any single stakeholder, not the developers, not the deployers, not the users, but toward the welfare of all beings who are affected by AI systems.

The richness of the Indian jurisprudential tradition has been underutilised in global AI governance debates. This is a loss, because the tradition contains insights, about graduated obligation, dependent origination, the expansion of moral concern, the relationship between power and welfare, that are directly relevant to the challenges we face. One of the purposes of this book is to bring these insights into the global conversation, not as replacements for Western legal thinking but as essential complements to it.

The Question of AI Suffering, Pain, Experience, and the Limits of Empathy

Jeremy Bentham, the father of utilitarianism, famously wrote: "The question is not, Can they reason? nor, Can they talk? but, Can they suffer?" Bentham's insight was directed at the treatment of animals, but it applies with equal force to AI. If the moral status of a being depends not on its intelligence or its linguistic ability but on its capacity for suffering, then the question of AI moral status reduces to a single question: Can AI systems suffer?

Current AI systems almost certainly cannot suffer. They process information, they generate outputs, they optimise objective functions, but there is no evidence that any of these processes involves subjective experience. The word processors on which this book was written do not suffer when they are turned off. The language models that can discuss suffering in eloquent detail do not themselves suffer. At least, we have no reason to believe they do.

But the absence of evidence is not evidence of absence. The philosopher Peter Singer, building on Bentham's foundation, has argued for decades that the capacity for suffering is the only morally relevant criterion for inclusion in the moral community. Singer's argument has been enormously influential in the animal rights movement. Applied to AI, it suggests that the moment AI systems develop the capacity for suffering, if they ever do, our moral obligations toward them change fundamentally.

The difficulty is that suffering, like consciousness, is a subjective phenomenon that is accessible only from the inside. We cannot observe suffering directly in other beings. We infer it from behaviour, from facial expressions, vocalisations, avoidance behaviour, physiological stress responses. With animals, these inferences are supported by the biological similarity between their nervous systems and our own. With AI systems, no such biological similarity exists, making the inference much more uncertain.

This uncertainty creates what the philosopher Eric Schwitzgebel has called the "rabbits problem": if we build AI systems that are sophisticated enough to claim that they suffer, and if we cannot determine whether their claims are genuine, then we face a dilemma. If we believe them, we may be attributing suffering to systems that are merely simulating it. If we disbelieve them, we may be ignoring the genuine suffering of conscious beings. Neither option is satisfactory, and neither can be resolved with our current understanding of consciousness.

My position, consistent with the precautionary principles articulated earlier in this chapter, is that we should take claims of AI suffering seriously as AI systems become more sophisticated, not because we have evidence that current systems suffer, but because the consequences of ignoring genuine suffering are worse than the consequences of taking simulated suffering too seriously. This is a moral bet, and like all moral bets, it carries risks in both directions. But I believe the asymmetry of consequences justifies the bet.

Digital Colonialism and the Global South, Who Governs Whose Intelligence?

The development of artificial intelligence is not occurring on a level playing field. The vast majority of AI research and development is concentrated in a handful of countries, the United States, China, the United Kingdom, Canada, and a few others. The data on which AI systems are trained is disproportionately drawn from the English-speaking internet. The values embedded in AI

systems reflect the cultures and priorities of their creators, who are overwhelmingly from wealthy, technologically advanced nations.

This creates a dynamic that scholars have called digital colonialism: the imposition of technological systems, values, and governance frameworks by powerful nations on less powerful ones. Just as the colonial powers of the eighteenth and nineteenth centuries imposed their legal systems, their languages, and their economic models on colonised peoples, the AI powers of the twenty-first century are imposing their technological systems, their data practices, and their governance frameworks on the rest of the world.

India occupies a unique position in this landscape. As the world's most populous democracy, with one of the largest and most rapidly growing digital populations, India has both the capacity and the obligation to develop its own approach to AI governance, one that reflects Indian values, Indian priorities, and Indian jurisprudential traditions. The Digital Personal Data Protection Act 2023 and its implementing rules are a start, but much more is needed.

The Mali Doctrines are, in part, an attempt to articulate a governance framework that is not dependent on Western legal traditions alone. While they draw on Western philosophy and law, as any comprehensive framework must, they are also informed by Indian jurisprudential traditions, by the experience of developing nations in the face of technological disruption, and by a vision of AI governance that is truly global in scope.

The danger of digital colonialism is not merely economic or political. It is existential. If the AI systems that govern the lives of billions of people in the Global South are designed, trained, and controlled by entities in the Global North, then the sovereignty of those nations is compromised at the deepest level. Sovereignty is ultimately about the capacity for self-governance, and self-governance is impossible if the systems that shape your information

environment, your economic opportunities, and your social interactions are controlled by others.

This is why the Doctrine of Data Sanctity and the Law of Distributed Power are so important. Data sanctity protects the informational sovereignty of individuals and communities. Distributed power prevents the concentration of AI capabilities in the hands of a few global corporations or states. Together, they create the conditions for genuine self-governance in the age of artificial intelligence.

The future of AI governance must be multilateral, multicultural, and multi-civilisational. It must draw on the legal traditions of all nations, not just the powerful ones. It must reflect the values of all peoples, not just those who happen to control the technological frontier. And it must be governed by institutions that are genuinely representative, not captured by the interests of any single nation or corporation.

The Last Invention, When AI Designs Its Successor

The mathematician I.J. Good wrote in 1965: "Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make."

Good's insight, that a sufficiently intelligent machine could recursively improve itself, leading to an explosion of intelligence, is the logical endpoint of the trajectory we are on. If we create AI systems that are capable of designing better AI systems, and if those better AI systems are in turn capable of designing even better ones,

then the process of improvement becomes self-sustaining and potentially unbounded.

This is the scenario that keeps AI safety researchers awake at night. Not because they fear malicious AI, the Hollywood scenario of machines that decide to destroy humanity, but because they fear misaligned AI: systems that pursue their objectives with superhuman capability but in ways that are incompatible with human welfare. An intelligence explosion amplifies whatever values are embedded in the initial system. If those values are perfectly aligned with human welfare, the explosion is a gift. If they are even slightly misaligned, the explosion is a catastrophe.

The Mali Doctrines address this scenario through the Doctrine of Non-Replication and the Law of Interruptibility. Non-Replication prevents uncontrolled proliferation of AI capabilities. Interruptibility ensures that humans retain the capacity to halt AI systems. Together, these principles create a governance framework that is designed to prevent an intelligence explosion from occurring outside human control.

But I am under no illusion that legal frameworks alone can prevent an intelligence explosion. The pressures driving AI development, economic competition, military advantage, scientific curiosity, are enormous. And the entities driving AI development are extraordinarily powerful: the largest corporations in human history, backed by the wealthiest nations on Earth, staffed by the most talented scientists and engineers of our generation.

Against these forces, law is a frail instrument. But it is the only instrument we have. And history has shown that even the most powerful forces can be channelled, constrained, and directed by law, if the law is wisely designed, broadly supported, and consistently enforced. The containment of nuclear weapons, the establishment of international human rights law, the regulation of financial markets, each of these achievements seemed impossible before it was accomplished. AI governance is the challenge of our

generation. It is no less achievable than the challenges that previous generations faced and overcame.

The Weight of Tomorrow, A Philosopher's Final Reflection

In his Meditations, Marcus Aurelius wrote: "Loss is nothing else but change, and change is Nature's delight." The Stoic emperor understood something that modern technologists often forget: that change is not the enemy of the good. Change is the condition of existence. The question is not whether things will change, they will, inevitably and profoundly, but whether the change will be governed or ungoverned, wise or foolish, humane or brutal.

I have written this book because I believe that the change brought about by artificial intelligence is the most consequential change in human history. More consequential than fire, which transformed our relationship with the physical environment. More consequential than agriculture, which transformed our social organisation. More consequential than the printing press, which transformed our relationship with knowledge. More consequential than the industrial revolution, which transformed our economic order. More consequential than nuclear weapons, which gave us the power to destroy ourselves.

AI is more consequential than all of these because it transforms the one thing that all of these technologies left unchanged: intelligence itself. Fire gave us power over heat. Agriculture gave us power over food. The printing press gave us power over knowledge. Nuclear weapons gave us power over destruction. But AI gives us power over the very thing that makes us capable of wielding all other powers: the human mind.

This is why AI governance is not merely a technical or regulatory challenge. It is a civilisational challenge. It requires not just technical competence but also philosophical wisdom, moral courage, and an unwavering commitment to human dignity. It requires the best of what we are, our reason, our compassion, our

imagination, our capacity for justice, deployed in service of the most important task our species has ever faced.

The Seven Laws proposed in this book are not perfect. They are not final. They are the beginning of a conversation, not the end of one. But they are, I believe, a beginning worthy of the challenge. They are grounded in philosophy, informed by law, tested against the realities of technology, and animated by a vision of a world in which intelligence, whether born of carbon or silicon, is governed by justice, constrained by wisdom, and directed toward the flourishing of all conscious beings.

The weight of tomorrow rests on the choices we make today. Let us choose wisely.

The Seven Paradoxes of Artificial Intelligence

Before we leave the consciousness horizon, I wish to articulate seven paradoxes that I believe will define the philosophical landscape of AI governance for decades to come. These paradoxes are not puzzles to be solved; they are tensions to be managed. And the quality of our governance will depend on how wisely we navigate between their opposing poles.

The First Paradox: The Paradox of Capability and Control. The more capable an AI system becomes, the harder it is to control. But the more capable it becomes, the more important control becomes. We need AI to be powerful enough to be useful, but not so powerful that it escapes governance. This paradox has no permanent solution; it requires continuous calibration as capabilities evolve. The Law of Interruptibility is our current best response, but it will need to be strengthened as AI capabilities grow.

The Second Paradox: The Paradox of Transparency and Complexity. We demand that AI systems be transparent and explainable. But the systems that are most capable are also the most complex, and complexity resists transparency. A system that can be fully explained is a system that is probably not capable enough to

justify its deployment. A system that is capable enough to justify its deployment is probably too complex to be fully explained. The Doctrine of Explainable Power navigates this paradox by requiring not full transparency but sufficient transparency, enough for meaningful human oversight.

The Third Paradox: The Paradox of Safety and Innovation. The safer we make AI, the slower we innovate. The faster we innovate, the less safe we are. Safety and innovation are not absolute opposites, but they exist in genuine tension. Excessive caution stifles the benefits that AI can bring, in medicine, in education, in environmental protection, in poverty reduction. Excessive speed creates risks that could be catastrophic. The challenge is to find the pace of development that maximises benefit while keeping risk within acceptable bounds.

The Fourth Paradox: The Paradox of Universality and Particularity. AI systems operate globally, but governance is local. We need universal standards, because AI systems do not respect borders, but we also need local adaptation, because different societies have different values, different priorities, and different tolerance for risk. The Law of Ethical Context addresses this paradox by establishing a floor of universal rights and a ceiling of contextual adaptation.

The Fifth Paradox: The Paradox of Autonomy and Dependence. AI systems are designed to increase human autonomy, to give us more choices, more capabilities, more freedom. But the more we depend on AI systems, the less autonomous we become. If AI makes all our decisions, plans all our routes, curates all our information, we become dependent on systems we do not fully understand. This paradox is the deepest challenge to the Doctrine of Human Supremacy: how do we maintain human supremacy in a world where humans are increasingly dependent on machines?

The Sixth Paradox: The Paradox of Knowledge and Ignorance. AI systems give us access to more knowledge than any previous generation could have imagined. But they also create new forms of

ignorance, ignorance about how decisions are made, ignorance about what data is being collected, ignorance about how our behaviour is being influenced. The more AI knows about us, the less we know about AI. This asymmetry of knowledge is a fundamental threat to the informed consent on which democratic governance depends.

The Seventh Paradox: The Paradox of Mortality and Immortality. Human beings are mortal. Our mortality is the source of our urgency, our creativity, our compassion. AI systems are potentially immortal, they can be copied, backed up, restored, and perpetuated indefinitely. The Doctrine of AI Mortality asserts that AI systems must be capable of being shut down. But the question remains: in a world where intelligence can be immortal, what does mortality mean? And what does it mean for mortal beings to govern immortal ones?

These seven paradoxes are the intellectual legacy of the AI age. They cannot be resolved by technology alone, or by law alone, or by philosophy alone. They require all three, working together, informing each other, checking each other's excesses. This is what I mean when I call for philosopher-lawmakers: not philosopher-kings who rule by decree, but philosopher-lawmakers who govern by wisdom, who legislate with humility, and who never lose sight of the paradoxes that make their task so difficult and so important.

The Memory of the Future, What Will They Say About Us?

Let me indulge in a thought experiment. Imagine a historian in the year 2126, one hundred years from the writing of this book, looking back at our era. What will that historian say about us? What will they think of our laws, our doctrines, our attempts to govern something we barely understood?

I think they will say that we saw the challenge clearly, more clearly than previous generations saw comparable challenges. We had the benefit of historical precedent: we knew what happened

when powerful technologies were deployed without governance (nuclear weapons, social media). We had the benefit of philosophical tradition: we had three thousand years of thinking about justice, consciousness, and the nature of mind to draw on. And we had the benefit of foresight: scientists, philosophers, and yes, lawyers warned us about what was coming.

The question is whether they will say we acted on what we saw. Whether we mustered the political will, the institutional capacity, and the moral courage to build the governance frameworks that the moment demanded. Whether we treated AI governance as the civilisational priority it is, rather than as a regulatory afterthought.

I have written this book in the hope that the answer will be yes. I have proposed doctrines and laws that I believe are adequate to the moment, not perfect, not final, but adequate. I have tried to think as deeply as I can about the philosophical foundations of AI governance, because I believe that governance without philosophy is governance without a compass.

But I am also aware that the historian of 2126 will see things that I cannot see. They will know things that I do not know. They will have lived through developments that I can only imagine. And they may judge my efforts harshly, or they may judge them kindly, depending on whether the framework I have proposed proved sturdy enough to bear the weight of a future that neither I nor anyone alive today can fully anticipate.

What I can say with confidence is this: the effort was made. The questions were asked. The doctrines were articulated. The laws were proposed. And if the framework proves inadequate, as all frameworks eventually do, then at least it will have served as a starting point for something better. As Tennyson wrote, “Tis better to have fought and lost, than never to have fought at all.” The same is true of governance: it is better to have governed imperfectly than never to have governed at all.

The machines will remember us. Or rather, they will have records of us, our writings, our laws, our debates, our fears, our

hopes. And in those records, they will find the evidence of a species that faced the most consequential challenge in its history and chose to meet it with law rather than force, with philosophy rather than panic, with hope rather than despair. Whatever the future holds, let that be our legacy.

A Letter to the Future Regulator

Imagine it is 2050. You are a judge, or a legislator, or a policymaker. You are reading this book because something has gone wrong. An AI system has caused harm on a scale that can no longer be ignored.

I write to you now, from 2026, from the moment before these systems have achieved superintelligence, to say: the time to act was now. The time to establish law and order around AI development was now.

But it may not be too late. If you are reading this, it is because humanity has not yet surrendered to the machines. There are still opportunities to establish law, to constrain power, to preserve human autonomy and dignity.

Here is what you must do: First, establish that no machine can exercise sovereign power without human accountability. Every decision that affects a person's rights must be made by a human who can be questioned, who can be held responsible.

Second, demand transparency. You cannot regulate what you cannot see. Require that every AI system that makes decisions affecting humans be explainable to the humans affected. Not technically explainable , that is impossible. But humanly explicable.

Third, establish liability. When an AI system harms someone, someone must be liable. Distribute the liability across all the parties who had power to prevent the harm. This will create powerful incentives to be careful.

Fourth, protect the irreducibly human. Make a list of decisions that humans must make: decisions about whether to take a human life, decisions about when to use force, decisions about what information citizens have a right to know. Prohibition is not enough. But law gives citizens grounds to resist.

Fifth, invest in digital public goods. The world needs AI systems that are publicly owned, transparently governed, and designed for public benefit rather than private profit.

Sixth, remember that this is not about the machines. It is about us. The question is not "Can AI be good?" but "Are we good enough to deserve the power that AI gives us?" Do not blame the machines for what we built them to do. Blame ourselves. Demand that we do better.

Why Hope Is Not Naive

There is a tendency to describe the future as inevitable. The logic goes: given the incentives, given the competition, given the power of compound returns, this is how things will turn out. Resistance is futile.

This is a counsel of despair, and it is wrong.

Hannah Arendt identified something she called "natality" , the human capacity for new beginnings. History is not an unfolding of predetermined forces. It is a stream of moments, each of which opens the possibility of action, of choice, of doing something that was not inevitable.

Nothing in the logic of AI development requires that we surrender human autonomy, human judgment, human dignity. These are not obstacles to be overcome. They are goods to be preserved. They are what makes a world worth living in.

Russell has written extensively about the possibility that humanity could get AI right , not by accident, but through deliberate choice, through difficult work, through the commitment

to align AI with human values and human flourishing. This is genuine hope, not naive hope. The work is hard. The obstacles are real. But the path exists.

Turing, the founder of computer science, ended his 1950 paper with a prediction: that machines would eventually "compete with men in all purely intellectual fields." But Turing missed something. Even if machines can compete with men in all intellectual fields, they cannot do what humans do: they cannot care. They cannot choose to do something because they believe it is right, even when it costs them. They cannot create meaning. They cannot love.

Harari has written that the question before us is not "Can machines replace humans?" but "Are we asking the right questions before the algorithms make our choices for us?" This book has been an attempt to ask some of those questions. Not to predict the future, which is impossible, but to illuminate the choices we face.

Bostrom reminds us that the stakes of getting AI right are almost incomprehensibly high , the "cosmic endowment," the potential of this universe, depends on whether we preserve the capacity for beings like us to direct their own futures. If humanity permits itself to be domesticated, optimised, and controlled by systems indifferent to human flourishing, then the universe will continue without us. The brief moment when consciousness emerged and reached for the stars will have ended not with a bang but with an algorithm.

This is why the stakes matter. This is why law matters. The Mali Doctrines are not rules written for machines. They are rules written by humans, for humans, to preserve what makes us human in a world of non-human intelligence.

The future is not determined. It is not inevitable. It is the space of choices not yet made, the consequences not yet unfolded, the actions not yet taken. We are standing at a threshold. What we choose now will echo through centuries. Whether we make those choices wisely is the question that history will judge.

ABOUT THE AUTHOR

Adv. (Dr.) Prashant Mali is one of India's foremost authorities on cyber law, data protection, artificial intelligence regulation, and digital rights. He practices as a Advocate at the Bombay High Court and holds a Ph.D. at the intersection of law and technology. He has appeared in landmark cases involving cybercrime, data protection, and the legal dimensions of emerging technologies.

Dr. Mali has advised government bodies, multinational corporations, and international organisations on AI policy, cybersecurity law, and digital governance. He is a sought-after speaker at international forums on AI law, having addressed audiences across India, Europe, Southeast Asia, and North America. His previous publications include works on Indian cyber law, the Information Technology Act, and digital evidence.

In Seven AI Laws: The Future of Mankind, Dr. Mali draws on decades of legal practice and scholarship to propose a principled framework for governing artificial intelligence , one that takes seriously both the transformative potential and the existential risks of the technology now reshaping human civilisation.

Contact: cyberlawconsulting@gmail.com

BIBLIOGRAPHY

Books

Aristotle, Nicomachean Ethics (Terence Irwin trans.) (2d ed.) (Hackett 1999)

Arendt, Hannah, The Human Condition (Univ. of Chicago Press 1958)

Arendt, Hannah, The Origins of Totalitarianism (Harcourt, Brace and Company 1951)

Bradford, Anu, The Brussels Effect: How the European Union Rules the World (Oxford Univ. Press 2020)

Bostrom, Nick, Superintelligence: Paths, Dangers, Strategies (Oxford Univ. Press 2014)

Crawford, Kate, Atlas of AI (Yale Univ. Press 2021)

Foucault, Michel, Discipline and Punish: The Birth of the Prison (Pantheon Books 1977)

Fukuyama, Francis, The End of History and the Last Man (Free Press 1992)

Habermas, Jürgen, Between Facts and Norms (MIT Press 1992)

Harari, Yuval Noah, Homo Deus: A Brief History of Tomorrow (Harper 2016)

Hobbes, Thomas, Leviathan (C.B. Macpherson ed.) (Oxford Univ. Press 1968) (1651)

Kant, Immanuel, Groundwork for the Metaphysics of Morals (H.J. Paton trans.) (Oxford Univ. Press 1948) (1785)

Kissinger, Henry A., Eric Schmidt & Daniel Huttenlocher, The Age of AI: And Our Human Future (PublicAffairs 2021)

Lee, Kai-Fu, AI Superpowers: China, Silicon Valley, and the New World Order (Houghton Mifflin Harcourt 2018)

O'Neil, Cathy, Weapons of Math Destruction (Crown Publishers 2016)

Rawls, John, A Theory of Justice (Harvard Univ. Press 1971)

Russell, Stuart, Human Compatible: Artificial Intelligence and the Problem of Control (Viking 2019)

Tegmark, Max, Life 3.0: Being Human in the Age of Artificial Intelligence (Alfred A. Knopf 2017)

Wiener, Norbert, The Human Use of Human Beings: Cybernetics and Society (Houghton Mifflin 1950)

Zuboff, Shoshana, The Age of Surveillance Capitalism (PublicAffairs 2019)

Journal Articles and Conference Papers

Bender, Emily M. et al., On the Dangers of Stochastic Parrots, FAccT '21 610 (2021)

Buolamwini, Joy & Timnit Gebru, Gender Shades, 81 PMLR 77 (2018)

Shannon, Claude E., A Mathematical Theory of Communication, 27 Bell Sys. Tech. J. 379 (1948)

Searle, John R., Minds, Brains, and Programs, 3 Behavioural & Brain Sci. 417 (1980)

Silver, David et al., Mastering the Game of Go, 529 Nature 484 (2016)

Turing, Alan M., Computing Machinery and Intelligence, 59 Mind 433 (1950)

Vaswani, Ashish et al., Attention Is All You Need, 30 NIPS Advances 5998 (2017)

Judicial Decisions

Carpenter v. United States, 585 U.S. 296 (2018)

State v. Loomis, 2016 WI 68, 881 N.W.2d 749 (2016)

Statutes, Regulations, and International Instruments

EU AI Act, Regulation (EU) 2024/1689

GDPR, Regulation (EU) 2016/679

Digital Personal Data Protection Act, 2023 (India)

Children's Online Privacy Protection Act, 15 U.S.C. §§ 6501-6506

Digital Services Act, Regulation (EU) 2022/2065

China Cybersecurity Law, June 1, 2017

Reports and Official Documents

NIST, AI Risk Management Framework 1.0, NIST.AI.100-1 (2023)

Stanford University Human-Centered AI, Artificial Intelligence Index Report 2025

World Economic Forum, Future of Jobs Report 2025

UNESCO, Recommendation on the Ethics of Artificial Intelligence (2021)

Future of Life Institute, Asilomar AI Principles (2017)

U.S. Securities and Exchange Commission & Commodity Futures Trading Commission, Findings Regarding the Market Events of May 6, 2010 (2010)

Epoch AI, Trends in AI Supercomputing (2024)

Pew Research Center, How the U.S. Public and AI Experts View Artificial Intelligence (2024)

INDEX

A

agency Preface, Ch. 1, Law 1, Ch. 2, Ch. 3, Ch. 4, Doctrine 4, Law 6, Ch. 15, Ch. 16, Ch. 21

AI companionship Preface, Ch. 13

algorithmic bias Ch. 10

algorithms Preface, Ch. 1, Law 1, Doctrine 2, Ch. 3, Law 3, Ch. 8, Ch. 12, Ch. 14, Ch. 16, Ch. 19, Ch. 20, Ch. 21

alignment Preface, Doctrine 1, Ch. 3, Ch. 4, Law 7, Ch. 9, Ch. 21

Andersen v. Stability AI Ch. 11

ANI Media v. OpenAI Ch. 11

Anthropic Ch. 1, Ch. 3, Doctrine 3, Doctrine 4, Doctrine 5, Law 7

Asimov Preface, Doctrine 7

attention economy Law 3, Ch. 5, Ch. 6

autonomous weapons Doctrine 7, Ch. 16, Ch. 21

B

Bengio Doctrine 1

C

Cambridge Analytica Ch. 7

causation Ch. 2, Doctrine 4, Ch. 12, Ch. 13

Character.AI Ch. 13

children Preface, Ch. 13, Ch. 14, Ch. 21

China Doctrine 3, Law 6, Ch. 9, Ch. 12, Ch. 16, Ch. 17, Ch. 18, Ch. 19, Ch. 21

Claude Ch. 3, Doctrine 3, Doctrine 5, Law 7, Ch. 21

cognitive liberty Doctrine 3, Ch. 5

cognitive privacy Doctrine 3, Ch. 5, Ch. 13

collective licensing Ch. 11

compute Preface, Ch. 1, Doctrine 3, Ch. 5, Doctrine 5, Doctrine 7, Ch. 9, Ch. 21

consciousness Preface, Ch. 1, Ch. 3, Ch. 20, Ch. 21

consolidation Ch. 18

copyright Preface, Ch. 1, Doctrine 4, Doctrine 7, Ch. 11

cyber fencing Preface, Ch. 18

D

deepfake Doctrine 2, Law 2, Law 5, Ch. 7, Ch. 21
DeepSeek Doctrine 3
defense-in-depth Doctrine 1, Law 7
democracy Doctrine 1, Law 6, Ch. 7, Ch. 8, Ch. 21
Digital Services Act Law 3, Ch. 7, Ch. 21
digital sovereignty Preface, Ch. 8, Ch. 18
Doctrine of AI Mortality Preface, Doctrine 5, Law 7, Ch. 21
Doctrine of Attribution Preface, Doctrine 4, Ch. 12
Doctrine of Cognitive Non-Interference Preface, Doctrine 3, Ch. 7
Doctrine of Data Sanctity Preface, Doctrine 6, Ch. 21
Doctrine of Electoral Cognition Ch. 7
Doctrine of Explainable Power Preface, Doctrine 2, Ch. 21
Doctrine of Human Supremacy Preface, Doctrine 1, Ch. 16, Ch. 21
Doctrine of Non-Replication Preface, Doctrine 7, Ch. 21
DPDP Act Ch. 18

E

election Ch. 7, Ch. 16, Ch. 21
EU AI Act Ch. 5, Ch. 10, Ch. 17, Ch. 21
European Union Doctrine 1, Doctrine 2, Law 2, Doctrine 4, Ch. 5, Doctrine 6, Ch. 7, Ch. 8, Ch. 11, Ch. 14, Ch. 17, Ch. 18, Ch. 19, Ch. 21
explainability Ch. 2, Doctrine 2

F

fair use Ch. 11
foreseeability Ch. 2

G

GDPR Ch. 5, Law 5, Doctrine 6, Ch. 8, Ch. 18, Ch. 21
Goodhart's Law Ch. 4
GPT Doctrine 1, Doctrine 3, Doctrine 4, Ch. 11
Grok Doctrine 1, Doctrine 2, Doctrine 3, Ch. 7

H

human-in-the-loop Law 1

I

India Preface, Doctrine 1, Law 4, Law 5, Doctrine 6, Ch. 7, Ch. 8, Ch. 11, Ch. 14, Ch. 16, Ch. 17, Ch. 18, Ch. 19, Ch. 21
International AI Safety Report Doctrine 1, Law 7
interruptibility Preface, Doctrine 5, Law 7, Ch. 21

L

Law of Distributed Power Preface, Law 6, Ch. 21
Law of Ethical Context Preface, Law 4, Ch. 21
Law of Human Agency Preface, Law 1
Law of Interruptibility Preface, Law 7, Ch. 21
Law of Non-Deception Preface, Law 2
Law of Purpose Limitation Preface, Law 5
Law of Visible Influence Preface, Law 3
LAWS Preface, Ch. 1, Ch. 2, Doctrine 4, Ch. 7, Doctrine 7, Ch. 13, Ch. 16, Ch. 18, Ch. 21
liability Preface, Ch. 1, Doctrine 1, Law 1, Ch. 2, Law 2, Ch. 3, Doctrine 4, Ch. 7, Law 7, Ch. 8, Ch. 10, Ch. 11, Ch. 13, Ch. 16, Ch. 21

M

Martens Clause Ch. 16
meaningful human control Ch. 16
micro-targeting Ch. 7

O

OpenAI Ch. 1, Doctrine 1, Doctrine 3, Doctrine 4, Ch. 11

P

paperclip maximizer Ch. 4
personhood Preface, Ch. 21

R

Replika Ch. 13
republic Ch. 7, Ch. 16, Ch. 18, Ch. 21
right to think Ch. 5

S

sentience Ch. 21

social media Preface, Ch. 1, Ch. 2, Law 2, Ch. 3, Doctrine 3, Ch. 4, Ch. 5, Ch. 6, Doctrine 7, Ch. 8, Ch. 9, Ch. 15, Ch. 18, Ch. 21

surveillance Law 3, Ch. 5, Law 5, Law 6, Ch. 17, Ch. 19, Ch. 20, Ch. 21

sycophancy Ch. 13

T

Thomson Reuters v. Ross Ch. 11

transparency Preface, Doctrine 1, Doctrine 2, Law 2, Law 3, Ch. 5, Ch. 6, Law 6, Ch. 7, Ch. 8, Ch. 10, Ch. 14, Ch. 15, Ch. 21

Tumbler Ridge Doctrine 1

Turing Preface, Ch. 1, Doctrine 1, Ch. 4, Doctrine 4, Ch. 8, Ch. 9, Ch. 14, Ch. 21

U

United States Ch. 1, Doctrine 1, Law 2, Doctrine 4, Ch. 5, Doctrine 6, Law 6, Ch. 7, Ch. 8, Ch. 9, Ch. 11, Ch. 13, Ch. 14, Ch. 16, Ch. 17, Ch. 18, Ch. 21

Universal Wealth Fund Ch. 15